DOCUMENT RESUME

ED 091 444 95 TM 003 662

TITLE	Preparation of a Filmstrip Unit on Basic Measurement Principles. Final Report.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	National Center for Educational Research and Development (DHEW/OE), Washington, D.C.
BUREAU NO	BR-0-9050
PUB DATE	31 Oct 73
CONTRACT	OEC-0-70-4777
NOTE	93p.
AVAILABLE FROM	The filmstrip "Planning a Test" is available from Educational Testing Service, Princeton, New Jersey 08540
EDRS PRICE	MF-$0.75 HC-$4.20 PLUS POSTAGE
DESCRIPTORS	Criterion Referenced Tests; Filmstrips; Guides; Instructional Films; *Measurement; Questionnaires; *Test Construction; *Testing; Test Reliability

ABSTRACT

A filmstrip with associated audio track has been developed to cover the major planning steps in the development of a measurement instrument such as a test or questionnaire. The filmstrip addresses the following six questions: Why am I testing? What should I test? Whom am I testing? What kinds of questions should I use? How long should my test be? and How difficult should my test be? A set of supplementary materials cover the following issues: learning how to develop tests, obtaining information about tests, preparing a test plan, kinds of test questions advantages and disadvantages, reliability, and criterion-referenced tests. The supplementary materials provide an expanded treatment of some of the issues raised in the filmstrip, but their primary function is as a guide to appropriate literature for those individuals who are seeking an intensive treatment of topics related to their particular interest and needs. The target audience for the filmstrip and supplementary materials includes students in introductory educational psychology, measurement, and research training courses as well as teachers, administrators, or other educational personnel engaged in inservice training. No previous training in measurement is assumed. (Author)

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED 091444

Final Report

Project No. 0-9050
Contract No. OEC-0-70-4777

John J. Fremer Jr., Project Director
Educational Testing Service
Rosedale Road
Princeton, New Jersey 08540

TM 003 662

PREPARATION OF A FILMSTRIP UNIT ON BASIC MEASUREMENT PRINCIPLES

October 31, 1973

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

Office of Education

National Center for Educational Research and Development

ABSTRACT

A filmstrip with associated audio track has been developed to cover the major planning steps in the development of a measurement instrument such as a test or questionnaire. The filmstrip addresses the following six questions: why am I testing, what should I test, whom am I testing, what kinds of questions should I use, how long should my test be, and how difficult should my test be? A set of supplementary materials cover the following issues: learning how to develop tests, obtaining information about tests, preparing a test plan, kinds of test questions: advantages and disadvantages, reliability, and criterion-referenced tests. The supplementary material provide an expanded treatment of some of the issues raised in the filmstrip, but their primary function is as a guide to appropriate literature for those individuals who are seeking an intensive treatment of topics related to their particular interest and needs. The target audience for the filmstrip and supplementary materials includes students in introductory educational psychology, measurement and research training courses as well as teachers, administrators, or other educational personnel engaged in in-service training. No previous training in measurement is assumed.

Final Report

Project No. 0-9050

Contract No. OEC-0-70-4777

PREPARATION OF A FILMSTRIP UNIT ON BASIC MEASUREMENT PRINCIPLES

Filmstrip Title: Planning a Test

John J. Fremer Jr., Project Director

Educational Testing Service

Princeton, New Jersey 08540

October 31, 1973

The research reported herein was performed pursuant to a contract with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
National Center for Educational Research and Development

I. INTRODUCTION

A. STATEMENT OF THE PROBLEM

Individuals in many different educational specialties are engaged in evaluation and information gathering activities that make use of tests, but many of the people so occupied have very little background or training in fundamental measurement principles. They do attempt to design and develop instruments to determine student knowledge, skills, or attitudes or the effectiveness of materials or procedures, but they do so without confidence or highly developed skills. The single largest group of such evaluators are classroom teachers, many of whom have not had a single college course in test construction. If a college course was taken it may or may not have covered the essential principles that should guide any measurement activity. Even when a teacher has had training that was adequate, the passage of time may have eroded the learning that did take place. Classroom teachers, however, are not the only evaluators who have insufficient training to be effective measurement practitioners. Fred Kerlinger, in a discussion of the training of educational researchers, made the following observation:

> "Many research projects are wrecked on the rocks of measurement difficulties. That the educational researcher needs to know measurement theory, especially theories of reliability and validity, as well as how to construct measurement scales, seems hardly necessary to mention. Yet many educational researchers seem clearly not to be very sophisticated in measurement if we are to judge from the published literature. No, psychometrics can no longer be ignored in educational research."[1]

Clear understanding of measurement principles is equally important to the developer of course materials. In this era of accountability and individualization of instruction, the developer must incorporate measurement into his materials so that student progress can be efficiently monitored. Measurement knowledge is also

---

[1] Fred N. Kerlinger, The Doctoral Training of Research Specialists. Teachers College Record, 1968, 69, 477-483.

required in both the formative and summative evaluation phases of development. The more general applicability of measurement to the role of developer, though, rests in its value as a framework for approaching experience. The measurement processes of specifying, designing, and validating serve to focus attention on that which can be observed. In some instances careful attention to what could constitute an observable outcome will permit the identification of discrete easily measured behaviors. More often the highly valued educational outcomes will resist such fine analysis but the attempt will prove valuable. Knowledge of measurement concepts is also valuable to the educational administrators who are called upon to communicate to legislators, school boards, alumni, or parents, the outcomes of the work of researchers, developers, and evaluators.

Clearly, then, it would be useful to develop materials emphasizing basic measurement principles to assist in the instruction of those training for, or already engaged in, evaluation, development, or research.

B. WHERE TO START

As anyone who has conscientiously addressed the task of instrument development can attest, there are many components of the development process important enough to be the focus of instructional materials. In the original proposal for this project, the following topics were suggested as illustrative ones:

1. Developing Specifications for Measurement Instruments
2. Writing and Evaluating Objectively Scoreable Test Questions
3. Writing and Evaluating Free Response Questions
4. Reliability of Measurement Instruments
5. Validity of Measurement Instrumer
6. Item Analysis and Test Analysis

The topic listed first "Developing Specifications for Measurement Instruments" was chosen for final development because of its clear primacy in any measurement enterprise. (The title of the final product -- "Planning a Test" represents a simplification that was intended to better communicate the content of the materials.) Tinkelman (1971) speaks to the issue of the importance of test planning:

> "Careful initial planning makes it possible to avoid pitfalls, assures more efficient procedures, and results in a better end product. On the other hand, inadequate initial planning may impair seriously the quality and usefulness of the final test. For example, it may be found that the test items available for use are inappropriate, or that a solid statistical foundation for compiling the final test forms is lacking, or that it is not possible to publish the test with suitable norms or interpretive materials. At the very least, inattention to planning can lead to waste and to delay due to failure to coordinate properly the various phases of test construction."[2]

C. TARGET AUDIENCE

Based on the considerations outlined above, the target audience for the filmstrip was identified as persons engaged in or receiving training in evaluation. The audience would include students in introductory educational psychology, measurement, or research training courses at either the undergraduate or graduate level. It would also include teachers, administrators, or other educational personnel who were taking special in-service training courses or workshops.

---

[2]Tinkelman, Sherman N. Planning the Objective Test. In Robert L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D. C.: American on Education, 1971.

D. CHOICE OF MEDIUM

The filmstrip was chosen as the instructional mode since it appears to be the most efficient medium of communication -- perhaps because of its ability to sustain attention via frequent changes in the visual material. The opportunity for joint reinforcement of learning by simultaneous auditory and visual presentations, probably contributes to the favorable results obtained with filmstrips. While motion picture films can also achieve these effects, filmstrips offer considerable economies over films. Production costs for films of the type described herein, would be five times that of filmstrips, and subsequent reproduction costs would be two or three times as great. In addition, the technical problems associated with the personal handling of films or the scheduling problems associated with arranging for the help of audio-visual technicians, when such technicians are available, reduces the number of instructors who are willing to use films.

II PROCEDURES

The initial activity of the project was the development of a detailed outline of the possible content for a filmstrip on testing planning. In this connection, a review was conducted of literature on the needs for measurement knowledge of the target population, and teachers, evaluators, researchers, and developers were interviewed. The outline was reviewed by members of these same groups, some of whom were also engaged in the training of researchers, developers, and evaluators. One facet of this review was the tentative allocation of some aspects of subject-matter content to the filmstrip and others to the supplementary material that will accompany the filmstrip. This process of deciding where to place essential material continued throughout the project. Using the revised outline as a guide, five successive drafts of a filmstrip script and of supplementary material were developed and reviewed. As part of this activity suggestions for visuals were obtained and were also reviewed.

Two small-scale evaluations of a preliminary version of the filmstrip were then conducted, one at Trenton State College (N=12) with students who were enrolled in an initial graduate course in tests and measurements and the other at the University of Colorado at Boulder (N=9) with graduate students who were enrolled in an introductory research methods course. Each class listened to a tape of the script and viewed transparencies containing content being considered for slides. The students and the instructors for each course gave a detailed critique of a number of aspects of the text and slides, including overall interest level, terminology, relative clarity of sections, and value of the visual material.

For each tryout group an Evaluation Form was used to collect student reactions to this material. This Evaluation Form is included as Appendix A to this Final Report. Since the evaluation dealt with a preliminary version of the filmstrip material, detailed reporting of the results seems inappropriate. It may be worth noting, however, that even with this pilot set of materials 18 of the 21 students found it to be either "interesting" or "about average" whereas only three found it "boring". Similarly 11 of the 21 students judged the material either "much easier to understand" or "somewhat easier to understand" than a textbook presentation whereas five students felt it was "about the same" and five students thought it was "somewhat more difficult to understand."

Consideration was given to each of the comments made by students and instructors on specific aspects of the text and visual materials. Many of the changes made were in the direction of simplifying both the text and the visual materials. Efforts were also made to highlight particularly significant points both in the text and slides. In this connection, a summary and review section was added to the materials.

Another aspect of the preliminary tryouts was the use of a test of measurement issues related to test planning. This test was administered both before and after students were exposed to the materials. Since the project budget did not include provision for development of such a test, it was constructed primarily from available items in ETS owned tests. For this reason the test used is not included in this final report. The match of test content with preliminary filmstrip content was imperfect at best, but an analysis of specific item responses did provide some evidence of problems with the coverage of some aspects of filmstrip coverage. On the 20 item test, students improved their scores from an average

of 9.7 items correct on pretesting to 11.9 items correct on posttesting. This data is not presented as evidence for the effectiveness of the filmstrip as the nature of the test and the fact that materials were only in a preliminary stage makes them an inappropriate basis for such judgements.

After revisions to the filmstrip were made on the basis of the informal field trials, the script was placed in the hands of a professional film maker -- Visual Education Corporation. Visual Education Corporation carried out the following tasks:

1. Provided a critique of the script as received from ETS
2. Rewrote the audio script and developed visuals
3. Created a storyboard and presented to ETS a sample of the style to be used in the artwork
4. Prepared a taped version of script and a draft set of slides -- (Reviewed by ETS staff on two occasions and by Office of Education staff on one, i.e., December 20, 1971)
5. Revised script and visuals in consultation with ETS and prepared final draft tape and slides for review by ETS staff
6. Made further revisions to slides and script and presented slides and audio for approval
7. Delivered answer print of filmstrip and audio
8. Delivered final copies of filmstrip and audio

In addition to the ETS staff work on the filmstrip, slides and script that took place in collaboration with Visual Education Corporation, the supplementary materials were prepared in draft and subjected to review and revision by ETS staff.

On March 31, 1972 the supplementary materials in draft form were sent to Mrs. Doris Epstein, Contract Officer, Research Training Branch, National Center for Educational Research and Development. A letter of July 27, 1972 by Mrs. Epstein called for a simplification of the filmstrip supplementary materials and for additional material regarding the evaluation of the filmstrip.

In accordance with Mrs. Epstein's recommendations ETS staff have revised the supplementary materials for the filmstrip. Appendix B of this report is the revised set. The original version of the supplementary materials is included as Appendix C to the report. Some of the concepts involved in test planning particularly those relating to statistical issues have proved difficult to simplify. In general, though, the revised materials have a reduced vocabulary load and simpler language constructions. These revised materials are offered as a replacement for the original materials rather than as an alternative set to be used concurrently with less well-trained groups.

A copy of the script containing the text of the final filmscript is included as Appendix D to this final report.

No provision was made in the budget for this project for an evaluation of the final product. The filmstrip has been used, however, by individuals at the Upstate Medical Center of the State University of New York, Oral Roberts University, and Kansas State Teachers College. Copies of letters from these users are included as Appendix E of this final report.

III RESULTS

The final products of this project are the filmstrip entitled, "Planning a Test" and the supplementary materials that are available to accompany this filmstrip. The filmstrip has a running time of 25 minutes and addresses itself to the following six questions:

1. Why am I testing?
2. What should I test?
3. Whom am I testing?
4. What kinds of questions should I use?
5. How long should my test be?
6. How difficult should my test be?

An attempt was made to provide an overview of the test planning process and to call attention to a number of significant issues that a test planner and developer would need to consider. The visual material was selected both to sustain interest and to call attention to significant aspects of the text. Much of the content in the area of test planning is abstract and theoretical but it proved possible in most instances to select real world situations which exemplified the theoretical points being made in the text.

The supplementary materials that were prepared under this contract cover the following topics:

Learning How to Develop Tests

Obtaining Information About Tests

Preparing a Test Plan

Kinds of Test Questions: Advantages and Disadvantages

Reliability

Criterion-Referenced Tests

The supplementary materials and the filmstrip form an integrated package, and should serve to encourage the interested reader and viewer to pursue the issues that are raised that relate most directly to on-going projects or problems. The supplementary material contained numerous guides to source material. One of the components of the supplementary material, for example, is an extensive guide to the available literature about tests and assessment devices. The supplementary materials, however, do emphasize that a single exposure to the concepts contained in the filmstrip and supplementary materials is not enough to equip anyone to undertake the difficult task of measurement. Rather, it can only make them aware of some issues that they may not have considered and give them some assurance that there does exist material which might help them toward solutions toward their own practical problems.

IV CONCLUSION

As any materials developer must realize, the important conclusions about the materials created will come from the individuals who do or do not use them. Do they help the measurement practitioner gain greater understanding of the steps that need to be carried out to produce a useful measurement instrument? Do they provide the encouragement to go out and seek additional information? It is possible, however, to review the experience of preparing the filmstrip and associated supplementary materials, and to indicate what aspects of the preparation process seem to be particularly productive or unproductive.

Perhaps the single most useful feature of the developmental process followed in this project was that of visiting classrooms and administering trial versions of the materials to individuals representative of the target population. By observing these individuals as they listened and watched the draft materials, the project director and a cooperating colleague were able to discover the need for revision in the final filmstrip that might never have been determined by reviews

by experienced measurement experts.

Another feature of the developmental process that may be worthy of note, is the ability of the trained film maker to work effectively with technical material when he is given an opportunity to pose questions and receive information about troublesome points. Although the current project did not provide an occasion to test this hypothesis, it seemed to the project director that it would have been possible to present the film maker with appropriate background reading, to conduct a series of oral interviews regarding the test making process, and to start a series of film maker-produced drafts that would have led to a final script with much less investment of time by the measurement-trained project director.

V LIST OF APPENDICES

A. Evaluation Form - Preliminary Version of Filmstrip

B. Supplementary Materials - Revised and simplified version

C. Supplementary Materials - M rch 31, 1972 version,
(superseded by October 31, 1973 version)

D. Script of Filmstrip Test

E. Letters from Users of the Filmstrip

1. Dr. Henry Slotnick to Dr. John Fremer - July 20, 1973
2. Dr. James S. Waldron to Dr. John Fremer - July 20, 1973
3. Mr. William W. Jennigan to Dr. Donald E. Hood - September 19, 1973
4. Dr. Howard P. Schwartz to Dr. John Fremer - October 18, 1973

APPENDIX A - "Evaluation Form - Preliminary Version of Filmstrip,"
Final Report - "Preparation of a Filmstrip Unit on Basic Measurement Principles"
October 31, 1973, Project No. 0-9050, Contract No. OEC-0-70-4777

FILMSTRIP EVALUATION

Overall, I found this filmstrip to be:

(A) extremely interesting

(B) interesting

(C) about average

(D) boring

(E) very boring

Compared to reading this material from a textbook, presenting it in the filmstrip made it:

(A) much easier to understand

(B) somewhat easier to understand

(C) about the same

(D) somewhat more difficult to understand

(E) much more difficult to understand

Please write a brief statement which summarizes your reaction to the use of the "Gumby" cartoon character in the slides.

Were there any terms or words used in the filmstrip that were confusing to you or which you did not understand? _____ Yes _____ No

If yes, which ones?

Please rate the following topics by checking the appropriate boxes for each topic listed:

	Level of difficulty (Check one)			Future Usefulness (Check one)		
	Difficult	About right	Easy	Useful	Somewhat Useful	Not very Useful
1. Is this test necessary?						
2. How will using a test help me to make evaluative decisions?						
3. Where can I find information about tests?						
4. What are test specifications and how are they developed?						
4a. What should I test?						
4b. Whom am I testing?						
4c. What kinds of questions should I use?						
4d. How long should my test be?						
4e. How difficult should my test be?						
5. The total filmstrip						

As we go back through the slides during this evaluation, please rate each one on this scale by placing a check mark in the appropriate column.

	Very Useful	Useful	Average	Distracting	Very Distracting
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					
25					
26					
27					
28					
29					

On the lines below, indicate specific improvements that you would suggest for any of the topics.

1. Is this test necessary? _____

2. How will using a test help me to make evaluative decisions? _____

3. Where can I find information about tests? _____

4. What are test specifications and how are they developed? _____

4a. What should I test? _____

4b. Whom am I testing? _____

4c. What kinds of questions should I use? _____

4d. How long should my test be? _____

4e. How difficult should my test be? _____

If you marked one or more slides as very useful, can you explain WHY you marked them (it) that way?

If you marked one or more slides as very distracting, can you explain WHY you marked them (it) that way?

NCERD Reporting Form — Developmental Products

NOV 2 6 RECD

1. Name of Product	2. Laboratory or Center	3. Report Preparation
Filmstrip: Planning A Test Supplementary Materials	Educational Testing Service Princeton, New Jersey	*Date prepared* October 31, 1973 *Reviewed by* _____

4. Problem: *Description of the educational problem this product designed to solve.*

1. Individuals in many different educational specialties are engaged in evaluation and information gathering activities that make use of tests, but many of the people so occupied have very little background or training in fundamental measurement principles.

2. A critical area for any educational project that involves test development or selection is that of test planning. A careful, comprehensive review of the issues involved in test planning and of strategies for dealing with these issues will be a valuable tool for teachers, course and curriculum developers, and researchers in training.

3. Depending on the role that a test planner will be called on to play and the nature of the project, different topics will need to be emphasized. For some topics intensive study will be required so the product will need to refer the user to appropriate source materials.

5. Strategy: *The general strategy selected for the solution of the problem above.*

1. The major issues involved in planning a test were identified by measurement specialists, teachers, and educational researchers.

2. Preliminary content outlines were developed for a filmstrip and supplementary materials. These were reviewed and revised.

3. Preliminary audio and visual material was developed and piloted.

4. A revised filmstrip script and suggested visuals were prepared, a professional film company developed several versions of a script and draft visuals.

5. After reviews by ETS and NCERD staff a final filmstrip and supplementary materials have been prepared.

6. Release Date: *Approximate date product was (or will be) ready for release to next agency.*	7. Level of Development: *Characteristic level (or projected level) of development of product at time of release. Check one.*	8. Next Agency: *Agency to whom product was (or will be) released for further development/diffusion.*
On or about: January 1, 1974, depending on contract negotiations.	___ *Ready for critical review and for preparation for Field Test (i.e. prototype materials)* ___ *Ready for Field Test* ___ *Ready for publisher modification* X *Ready for general dissemination/diffusion*	Educational Testing Service has requested an opportunity to distribute and has preliminary approval.

10-71-A (D)

9. Product Description: *Describe the following; number each description.*

- *1. Characteristics of the product.*
- *2. How it works.*
- *3. What it is intended to do.*
- *4. Associated products, if any.*
- *5. Special conditions, time, training, equipment and/or other requirements for its use.*

1. Characteristics of the product - "Planning a Test" is a 25 minute filmstrip with an associated audio cassette and a set of printed supplementary materials.

2. How it works - The filmstrip and supplementary materials are likely to be most effective when used as the introduction to a class or in-service training session.

3. What it is intended to do? - "Planning a Test" provides and introduction to, and suggestions for further study of, the following topics or questions:

Filmstrip	Supplementary Materials
- Why am I testing?	- Learning how to develop tests
- What should I test?	- Obtaining information about tests
- When am I testing?	- Preparing a test plan
- What kinds of questions should I use?	- Kinds of questions: Advantages and Disadvantages
- How long should my test be?	- Reliability
- How difficult should my test be?	- Criterion-referenced tests

4. Associated products - See number one above

5. Requirements for Use -

- A filmstrip projector and screen with an associated cassette player or with an independent cassette player.

- A room to accomodate group for showing filmstrip.

- Reproduction equipment to prepare multiple copies of the supplementary material. Supplementary material is in a form suitable for photographic reproduction.

10. **Product Users:** *Those individuals or groups expected to use the product.*

Learners - (no previous training in measurement is assumed)
- students in introductory psychology, measurement, and research training courses.
- teachers, administrators, materials developers, or other educational personnel involved in in-service training programs

Teachers - Staff teaching groups mentioned above, prior measurement training would be highly desirable.

11. **Product Outcomes:** *The changes in user behavior, attitudes, efficiency, etc. resulting from product use, as supported by data. Please cite relevant support documents. If claims for the product are not yet supported by empirical evidence please so indicate.*

1. The final product has not been used in formal field trials.

2. Pilot work with a preliminary version of the filmstrip indicated that students maintain high interest level and a preference for the filmstrip over textbook treatments. Subjects were 21 students in measurement training courses.

3. Very favorable reports have been received from three college sites where the final filmstrip was used in a measurement course, a full day in-service workshop and a single training session.

4. The expected outcomes for this product are a significantly increased awareness of the complexity of an adequate test development effort. The viewer and listener should recognize the steps that need to be taken to produce adequate instruments.

5. Those learners who face specific test development tasks should be provided with a resource for pursuing in depth the issues that relate to their activity.

12. **Potential Educational Consequences:** *Discuss not only the theoretical (i.e. conceivable) implications of your product but also the more probable implications of your product, especially over the next decade.*

The job of improving competency in measurement is too big to be effected greatly by a single filmstrip. If it arouses user interest it will be able to reduce some of the most obvious measurement absurdities that are now being pursued in schools, colleges, research groups, and other places.

It will have the greatest impact when used by trained measurement people as an introductory item in a class or in-service training session.

13. Product Elements: *List the elements which constitute the product.*	**14. Origin:** *Circle the most appropriate letter.*
1. Filmstrip: "Planning a Test" 25 minutes	(D) M A
Reel of tape	(D) M A
Audio Cassette	(D) M A
2. Supplementary Materials for Filmstrip: "Planning a Test"	(D) M A
	D M A
	D M A
	D M A
	D M A
	D M A
	D M A
	D M A
	D M A
	D M A
	D M A
	D= Developed M= Modified A= Adopted

15. Start-up Costs: *Total expected costs to procure, install and initial use of the product.*

To be determined.

16. Operating Costs: *Projected costs for continuing use of product after initial adoption and installation (i.e., fees, consumable supplies, special staff, training, etc.).*

1. Instructor time for preparation and use of materials in training session.
2. Cost of further reproduction of supplementary materials for new users.

17. Likely Market: *What is the likely market for this product? Consider the size and type of the user group; number of possible substitute (competitor) products on the market; and the likely availability of funds to purchase product by (for) the product user group.*

1. Tests and Measurement classes at colleges and universities.
2. In-service training courses for teachers, materials developers, and educational researchers.
3. Continuing education programs.

APPENDIX B - "Supplementary Materials-Revised and simplified version,"
Final Report-"Preparation of a Filmstrip Unit on Basic Measurement Principles"
October 31, 1973, Project No. 0-9050, Contract No. OEC-0-70-4777

Supplementary Materials for Filmstrip -- "Planning a Test"[1]

by

John Fremer[2]

Educational Testing Service

The topics covered in this document are:

Learning How to Develop Tests
Obtaining Information About Tests
Preparing a Test Plan
Kinds of Test Questions: Advantages and Disadvantages
Reliability
Criterion-Referenced Tests

Learning How to Develop Tests

The filmstrip "Planning a Test" provides an overview of the test development process and calls attention to a number of significant issues for the test planner and developer. If you have already taken a course in Tests and Measurements, the filmstrip reviews some material familiar to you. If you are only now taking such a course or if you have no training in measurement, the filmstrip should make you aware that there is much to be learned about test development. If you want to learn more by reading on your own, you might

---

[1]The filmstrip "Planning a Test" and these supplementary materials were prepared by Educational Testing Service under Contract No. OEC-0-70-4777 (Project No. 0-9050) with the U. S. Department of Health, Education and Welfare, Office of Education, National Center for Educational Research and Development. Single copies of supplementary materials in a form suitable for photographic reproduction will accompany each filmstrip order. Address for orders: Educational Testing Service, Princeton, New Jersey 08540. These supplementary materials may be reproduced without permission from the Office of Education or Educational Testing Service.

[2]Substantial contributions to the development of these supplementary materials were made by Clair Bowman, Miriam Bryan, Eleanor Horne, S. Donald Melville, and Michael Zieky.

obtain a good elementary measurement textbook. There are a number available including:

Cronbach, L.J., Essentials of Psychological Testing (3rd ed.) New York: Harper & Row, 1970.

Thorndike, R. L., and Hagen, E., Measurement and Evaluation in Psychology and Education (3rd ed.) New York: Wiley, 1969.

If you want detailed information on particular topics, the best single source of authoritative articles and references is:

Thorndike, R. L. (Ed.) Educational Measurement. Washington: American Council on Education, 1971.

In addition to an introductory chapter by the editor Robert L. Thorndike, the major areas covered by this book are:

Test Design, Construction, Administration, and Processing (7 chapters)
Special Types of Tests (3 chapters)
Measurement Theory (5 chapters)
Application of Tests to Educational Problems (4 chapters)

For a more complete list of textbooks and other source and reference materials in the area of educational measurement, you can write Educational Testing Service, Princeton, New Jersey 08540, for the following pamphlet:

Locating Information on Educational Measurement: Sources and References. (2nd ed.) Princeton: Educational Testing Service, 1969.

It is also possible to obtain from The Psychological Corporation, 304 East 45th Street, New York, New York 10017, a list of documents on testing issues entitled the Test Service Bulletins. Harcourt, Brace, Jovanovich, Inc., 757 Third Avenue, New York, New York 10017, can provide a list of Test Service Notebooks that discuss test theory, administration of testing programs, proper use of test results, and results of research studies.

Obtaining Information About Tests

A great deal of information about tests is available. However, it is often hard for an individual beginning a test development project to find just what he needs. Finding the time can be a serious problem. Finding the appropriate materials may also be difficult as not all libraries contain large collections of materials about tests and testing.

There is no one ideal way to search for background information, but some steps can be identified that will usually be helpful. They are the following:

I. Determining your needs.

II. Survey of information about existing tests.

III. Literature search.

IV. Test collections, other sources.

I. Determining Your Needs

No search for information about tests is likely to be very productive unless you spend a good deal of time first deciding what it is that you are looking for. Some important points to consider are:

A. What are your testing objectives?

1. How do they relate to the overall objectives of your project?
2. What type of information are you seeking?
3. How will you use it?
4. What decisions do you have to make?
5. How can information from tests help you to make these decisions?

B. What are the limits to your search and test development project?

1. Time and Schedule -- When do you need to have your test or questionnaire? How much time can you devote to your search for information and to the task of test development?

2. Help -- Are you on your own or do you have other people working with you? If you have help, how should you divide responsibility?

3. Money -- Can you afford to buy materials about tests or to design tests that require expensive equipment?

4. Whom will you be testing? -- Consider such factors as age, previous familiarity with tests, language development, and motivation to cooperate.

5. Available Facilities -- What search sources, such as a college library, are convenient? What sources are harder to reach but still possible?

C. Keep careful records of your search. Design a worksheet or checklist for recording information that you can use to compare or select measures.

II. Survey of Information About Existing Tests

Your own search will be greatly assisted if you can locate a recent book or article that reviews tests in your area of interest. You can identify for further study the references that relate most closely to your interests, and you can usually determine what publications would be likely to publish additional articles relevant to your project. The following references will often be useful:

A. Mental Measurements Yearbook Series (Gryphon Press, Highland Park, New Jersey)

The volumes in this series include description of tests, critical reviews, publishers' directories, and bibliographical references.

1. Mental Measurements Yearbooks (MMY)
2. Tests in Print
3. Reading Tests and Reviews
4. Personality Tests and Reviews

B. CSE: Elementary School Test Evaluations and CSE--ECRC Preschool/ Kindergarten Test Evaluations

These volumes include ratings of tests on a number of criteria. They are published by the Center for the Study of Evaluation and the Early Childhood Research Center, UCLA Graduate School of Education, Los Angeles, California.

C. NCME Measurement News

This newsletter of the National Council on Measurement in Education contains general articles on testing issues as well as announcements of new tests and lists of test reviews.

D. Test Collection Bulletin (TCB) -- ETS, Princeton, New Jersey

This is a quarterly digest of information on tests and services which generally have become available after the publication of the most recent Mental Measurement Yearbook. It describes both commercially available tests and tests used experimentally. The Bulletin does not evaluate the tests listed, but it does provide references to test reviews.

E. Promotional Materials from Test Publishers

Check publishers' catalogs and announcements for references to tests, services, and technical data on specific measures made available after the publication of the Mental Measurements Yearbook.

F. Other Appropriate Reference Materials

"An Annotated Bibliography of References to Tests and Assessment Devices" may be obtained from the ETS Test Collection, Educational Testing Service, Princeton, New Jersey 08540.

III. Literature Search

You may be able to locate articles on tests in your area of interest by using reference sources that provide abstracts or by using the ERIC system. You can also search the most recent issues of professional testing and research journals.

A. Traditional Reference Tools

These sources of abstracts of educational and psychological literature can usually be found on the reference shelves of college libraries.

1. Psychological Abstracts
2. Education Index
3. Research Studies in Education
4. Dissertation Abstracts

B. Educational Resources Information Center (ERIC)

ERIC is actually a series of clearinghouses, each specializing in a specific area of education. Reports and articles are collected and indexed in ERIC publications: (1) Research in Education, (2) Current Index to Journals in Education, and (3) Clearinghouse Publications (frequently a source of state-of-the-art papers). Items cataloged in the ERIC system can be ordered in microfiche or hard copy. The ERIC Clearinghouse on Tests, Measurements and Evaluation is located at Educational Testing Service, Princeton, New Jersey.

C. Use of Information Directories

Particularly useful is the Encyclopedia of Information Systems and Services, edited by Anthony Kruzas. (New York: Edward Brothers, Inc., 1971)

D. Professional Journals

Testing and educational research journals not only provide reports of developmental and research activities, but also related theoretical papers. They often provide test reviews which may also be referenced in the Test Collection Bulletin and the NCME Measurement News. The advertisements in the journals are an excellent source of information on new materials and services available from commercial publishers and research organizations.

E. Textbooks in the Area Under Study

Are tests included, described, or mentioned? Scan the bibliographies for additional sources of information.

IV. Test Collections, Other Sources

A. Educational Testing Service Test Collection (Princeton, N. J.)

B. Head Start Test Collection (ETS, Princeton, New Jersey -- Funded by the Office of Child Development, Department of Health, Education and Welfare)

Both of these test collections provide on-site, telephone, and mail reference services. You can write for lists of publications that summarize available instruments in specified subject or skill areas.

C. Research and Development Centers and Child Study Laboratories, etc.

D. Professional Organizations and Special Interest Groups such as:

American Printing House for the Blind
International Reading Association
American Association for Health, Physical Education and Recreation

Preparing a Test Plan

Even before searching for information about tests, it is wise to give careful thought to your objectives for testing and to develop a preliminary test plan. After you have exhausted the available sources of information about tests, however, you may decide to develop your own tests. You will then need more detailed and structured plans. An extremely useful source of

guidance in this connection can be the chapter entitled, "Planning the Objective Test" by Sherman Tinkelman in Educational Measurement, edited by R. L. Thorndike. Other sources of information on preparing tests include:

Ebel, R. L., Measuring Educational Achievement. Englewood Cliffs, New Jersey: Prentice-Hall, 1965.

This text deals primarily with achievement tests prepared by teachers and professors for use in their own classes. It emphasizes methods of developing your own tests and evaluating the items, rather than the selection and use of standardized tests. Included are practical suggestions for planning, constructing, administering, scoring, and analyzing the results of classroom tests. No previous special training in educational measurement is assumed.

Educational Testing Service, Making the Classroom Test: A Guide for Teachers. (3rd ed.) Princeton: Educational Testing Service, 1973.

This pamphlet reviews the plans and procedures used by four hypothetical teachers to prepare good tests. It considers a number of special problems faced in the writing and scoring of tests.

If you are primarily concerned with the measurement of attitudes and interests rather than with measuring skills or knowledge, the following books can be useful:

Edwards, Allen L., Techniques of Attitude Scale Construction. New York: Appleton-Century-Crofts, 1957.

The author notes that his book, "...is intended for those who may desire to measure attitudes toward something in which they are interested, but who fail to find an appropriate scale available."

Robinson, John P., and Shaver, Phillip R., Measures of Social Psychological Attitudes. Ann Arbor, Michigan: Institute for Social Research, The University of Michigan, 1969.

A review and evaluation of 112 scales for measuring attitudes such as: life satisfaction and happiness, self-esteem, dogmatism, sociopolitical attitude, social values, attitudes toward people, and religious attitudes. A copy of each scale is included.

Shaw, M. E., and Wright, J. M., Scales for the Measurement of Attitudes New York: McGraw-Hill, 1967. 604 pp.

Not addressed to a particular audience, this book brings together a number of useful scales suitable for research purposes and group testing. The authors caution against the use of these scales for individual measurement, diagnosis, or personnel selection. Topics include the nature of attitudes and methods of scale construction; the scales, presented in eight chapters; and evaluation and suggestions for improvement.

One of the major parts of test planning is identifying the content for the test. You need to obtain a representative and balanced sample of tasks or questions. When developing achievement tests in subject-matter areas, it will be useful to consider questions such as the following:

What are the important things you would expect a person who has studied this subject to know?

What intellectual skills should he have acquired?

What level of understanding of the material should he be required to demonstrate?

What is the relative importance of these various elements?

Whenever possible, you should analyze the materials that were used in instruction in the particular subject-matter area. Among the sources of information that might be used are:

Textbooks and teachers' manuals -- When you prepare a test for your own course, you will want to consider just what assignments were particularly stressed so that this material can be emphasized in the test. When preparing tests for courses that you have not yourself taught, as would be the case in departmental testing or in a research setting, it will be necessary not only to review assignment sheets but to determine precisely what sections of the material were actually covered.

Course syllabi, curriculum guides, lesson plans, lecture notes, laboratory activities, films, and filmstrips -- Again for most purposes, you need to differentiate between what should have been covered and what actually was covered. Only in the unusual situation where you are addressing the evaluation of a course as a type of outside auditor who can only "go by the book" should you proceed as though a pre-course plan has actually been followed.

The content categories that you develop should be comprehensive ones covering all areas of interest. It will be useful to group the material into meaningful clusters and to determine the emphasis to be given each cluster in the final test.

In addition to identifying the content to be covered by a test, you will also need to ask what level of skill or understanding the student should demonstrate. Each of the content areas will contain not only factual material but material requiring inferences and the analysis of relationships. When considering the different skills or abilities to be tested, it will be useful to start by reviewing the appropriate one of the following taxonomies of educational objectives:

Bloom, B. S. (Ed.), Taxonomy of Educational Objectives: The Classification of Educational Goals; Handbook I: Cognitive Domain. New York: David McKay Co., Inc., 1956. 207 pp.

This volume is for educators and research workers who deal with curriculum and evaluation problems. It classifies the cognitive goals in education -- those goals primarily involving intellectual considerations. Part I explains the nature and development of the taxonomy, and it describes the principles and problems in classifying educational objectives. Part II presents the hierarchical taxonomy and illustrative materials for each level: knowledge, comprehension, application, analysis, synthesis, and evaluation.

Krathwohl, D. R., Bloom, B. S., and Masia, B. B., Taxonomy of Educational Objectives: The Classification of Educational Goals; Handbook II: Affective Domain.

This second book in the Taxonomy series is devoted to the affective goals of education -- those primarily concerning emotional or feeling behaviors of students such as appreciation, attitudes, and values. Part I describes the nature of the affective domain and the classification structure, and it describes the evaluation of affective objectives at each level of the structure. It also analyzes the relation of the affective to the cognitive domain.

The objectives for a course or the variables in research may come from the cognitive or the affective or even a third domain, the psychomotor. A review of only the cognitive domain is carried out here, however, since the same principles apply to any of the areas. It was for this reason that the

filmstrip used the cognitive skills dimension as one of the two dimensions of the content-skill matrix.

As noted in the filmstrip, the cognitive domain can be subdivided into six categories: knowledge, comprehension, application, analysis, synthesis, and evaluation.

The knowledge level may be equated with recall of information. It may range from recall of specific bits of information to the recall of universals and abstractions in a field.

Comprehension incorporates most of what teachers mean by "understanding." If an individual comprehends, he knows what is being communicated by a stimulus. However, he is not necessarily aware of either the implications of the stimulus or how it relates to other material.

Application involves using what has been learned. If an individual uses an abstraction learned in one setting in another setting, he is applying his knowledge.

An individual is operating at the analysis level when he breaks down a stimulus into component parts. He may do this to indicate how it is organized, what is conveyed, how it is conveyed, etc.

Synthesis is the reverse of analysis. Here the individual puts elements together to form a unified whole that did not exist prior to his efforts.

Evaluation, as the word clearly implies, is making judgments of value. It involves judgment of the extent to which stimuli satisfy criteria.

These six categories appear to be logically hierarchical in nature, with evaluation requiring all of the other skills, synthesis requiring all but evaluation, and so forth. Empirical evidence with which to either verify or reject the logical evidence is largely lacking. This does not detract greatly from the utility of the classification system. It should suggest caution in its use, however. One further note is in order. For any given individual, the category system is designed to indicate only what the test taker should be doing to respond to any given task. What the teacher or researcher intends and what the individual is actually doing may be quite different. There are obvious examples. When an individual has already had an experience in which he has had to evaluate the worth of a particular argument, to have him re-evaluate the same argument against the same criteria a second time is merely an exercise in recall of information -- knowledge level behavior. He must be using a different criterion or evaluating a different argument for his behavior to be evaluative.

The volumes containing descriptions of cognitive and affective domains provide examples of objectives at each level and test questions related to these levels.

Kinds of Test Questions: Advantages and Disadvantages

A number of considerations will influence the selection of questions for a test. Obviously, you will want to measure the kinds of knowledge or skills that you are particularly interested in. To some extent, the nature of the subject matter will influence your choices as the following examples illustrate:

1. If you want to know how a student feels about some issue, you will often find it efficient to ask his degree of agreement with statements relating to that issue.
2. If you are interested in the student's ability to translate from another language into English, your questions will have to include material in that other language.
3. If you are interested in evaluating a student's map-reading skills, the questions should use maps.

In some situations you may want to consider the use of essay questions, because they require a student to express answers in his own words without the benefit of suggested possibilities. Essay tests can be prepared quickly as there are few questions to write. If necessary, the questions can be written on a classroom chalkboard. Moreover, the use of essay questions very largely eliminates guessing. Essay questions can serve to measure some higher level abilities when pupils are required to present evidence, evaluate, analyze, solve new problems, or approach problems in a new way. Unfortunately, too many so-called essay questions do <u>not</u> do that. These may be questions such as: "Name the six largest cities in the United States" or "List the characters in each of the following stories." Essay questions may be ambiguously stated or be so undirected and general that a pupil can bluff or "talk around the subject." Poor essay questions are easy to write. Writing more challenging essays requires considerable thought.

The principle disadvantage of the essay question is, of course, the unreliability of the scoring of the answer. Why is it so difficult to achieve reliability in scoring answers to essay questions? One reason is

that the criteria used to make judgments differ. One grader may think that an answer is good; another may think that it is poor. This may not be the result of the inadequacy of one of the graders; it may be the result of an honest difference of opinion on the relative merits of an answer. In the classroom setting, some teachers believe that the pupil's ability to express himself must be taken into account; these teachers deduct credit on answers to essay questions in social studies, science, and other subject-matter areas for poor English expression. Other teachers believe that a pupil should not be penalized if he knows the subject-matter but cannot express himself especially well.

A variety of irrelevant factors may affect scoring results. If the essay question is concerned with a controversial topic, the grader's judgment is likely to be influenced by his own convictions. Then, too, the teacher's judgment may be influenced by how a paper looks -- the easier the paper is to read, the higher is the score assigned. Or the teacher's judgment may be influenced by the "halo" effect. There may be a tendency to mark in terms of work in class or even in terms of what is expected of the class to which the pupil is assigned. The "halo" effect may also operate from question to question, the quality of the answer to the first question influencing the scoring of the answers to subsequent questions. Add to all of these sources of unreliability of scoring the fact that teachers frequently correct tests after a long school day, even late in the evening, and that their scoring is likely to be somewhat erratic as a result of real fatigue. Sometimes teachers who rescore essay questions after an interval of time find themselves coming up with quite different scores.

Beyond the problem of reliable scoring, there are two other limitations to essay questions which should be mentioned. First, there is the possibility of inadequacy of sampling when only a few questions are selected to cover a large content area. Then, there is the large penalty per question that will result if the pupil does not know the answer to the question or, even more serious, if he knows the answer but does not understand the question.

Finally, a choice of essay questions is frequently permitted. If the performance of pupils is to be compared one with another, a choice among essay questions should not be permitted. Without elaborate equating of the questions on the basis of their difficulty for the pupils who attempt them, there is no way of knowing, for example, how well the pupils who chose to answer questions 1, 2, and 4 would have done had they chosen to answer questions 3, 5, and 6 instead. And the better pupils who attempt the more challenging questions sometimes write less acceptable answers than do the less talented pupils who are satisfied to answer the easier questions.

To sum up the discussion of essays, although they are appropriate in some circumstances, they tend to be an inefficient method of obtaining information about individuals. They are useful as measurement devices only when multiple, independent grading is arranged, using graders who can agree in advance on common criteria. For a comprehensive discussion of essay examinations, see the chapter by Coffman in Educational Measurement.

Now what about questions of the objective type? They have at least four advantages:

1. They permit wider sampling of learning in a relatively short time, making spotty preparation on the part of the pupil more obvious.

2. They can be reliably scored. If the items are unambiguous and the test has been keyed properly, the scoring errors will be clerical errors rather than errors of judgment.

3. They are more easily scored. Scoring time is reduced. The teacher is freed from suspicion of partiality. Frequently, they can be scored by the pupils themselves.

4. They lend themselves rather readily to item analysis. The teacher can, over a period of time, assemble a file of questions, retaining from each test the questions that are of proper difficulty, and that discriminate well between high and low achievers.

One limitation of questions of the objective type is that they are difficult to construct if they are to test anything more than memory. All types of objective questions, especially multiple-choice questions, can test for recognition of assumptions, interpretation of data, recognition of limitations, application of principles, and a variety of other higher intellectual abilities and skills. Experience with several types of objective questions, however, suggests that some are more efficient for such purposes than others. Because the construction of objective questions that do measure understanding and thinking requires a large amount of time and considerable ingenuity, teachers are likely to be content with questions testing principally knowledge of facts -- sometimes very trivial facts. This limitation of questions of the objective type, then, is not a limitation inherent in the questions themselves but a limitation imposed by the test constructor.

Five types of objective questions are popular with classroom teachers -- completion, true-false, matching, classification, and multiple-choice. All of these types can serve useful purposes if the questions are well drawn and the answers to be supplied or chosen are definite and specific. They are all easily adapted to classroom situations and a large number of them can be administered in a relatively short time. In most testing situations, a balanced blend of several types of objective questions will prove most effective. When more than one type is used, however, it is a good idea to group together questions of the same type so that the pupils will not be confronted with too frequent changes in directions.

The textbooks already mentioned (Cronbach, Ebel, and Thorndike and Hagen) provide guidance on types of questions and how to write them. In addition, you may want to look at the following:

Gerberich, J. R., Specimen Objective Test Items: A Guide to Achievement Test Construction. New York: Longmans, Green and Co., 1956.

This book is designed primarily for teachers. It serves the test specialist by providing several systematic classifications of instruments and techniques used in achievement testing. The coverage of the book is as follows: measurement of educational achievement (1 chapter); specimen objective achievement test items (11 chapters); classification of objective achievement test items (3 chapters); and tests, nontest tools, and techniques used in achievement measurement (2 chapters).

Sanders, N. M., Classroom Questions: What Kinds?
New York: Harper and Row, 1966.

Intended for teachers, particularly social studies teachers, and other makers of classroom test questions, this book is based on Bloom's Taxonomy of Educational Objectives. Chapter titles include: "Questions Designed for More Than Memory"; "Memory"; "Translation"; "Interpretation"; "Application"; "Analysis"; "Synthesis"; "Evaluation"; and "Planning for Questioning."

Reliability

Reliability is a term for the dependability of a measurement. If we could measure the same set of people again and again with the same or comparable instruments or procedures, would we get the same or similar results on each testing? The answer is "no." Measurement, like any other human endeavor, involves a certain amount of error. The errors can be of two types: there is a systematic error, such as when scores on an instrument are all biased in one direction. Examples of biased instruments would be a ruler that is too short, a scale that adds five pounds to each weighing, a test that has a miskeyed question. Scores on such defective instruments may be wrong, but they will be wrong consistently, dependably, predictably, and reliably. Random errors, however, as the name implies, occur by chance. They are temporary and shifting, often due to unknown causes such as fluctuations in mood, health, motivation, and so on. The greater the amount of random error in a measuring instrument, the less the reliability of that instrument.

To help us think about reliability, a good example would be that of two watches -- one very fine and expensive, the other one rather cheap and poorly made. The fine watch will measure time reliably and dependably. Even if it is set to the wrong time so that it does not give an accurate estimate of the correct time, it will still be consistent, or, in this sense, "reliable." The cheap watch, on the other hand, will have a certain amount of purely random error. Sometimes it will be too fast, sometimes it will be too slow, and it may be impossible to predict which way the error is directed. It is this kind of random, unpredictable error that makes the cheap watch unreliable.

One very good definition of reliability is that reliability is a proportion. It is the proportion of the score that is a true measurement of the thing being measured. Obviously, the more random error, the smaller the proportion of truth in a measurement. The less truth in the measurement -- the smaller the proportion of truth to error -- the lower reliability will be.

Unfortunately, there is no direct way of determining reliability by discovering what proportion of a measuring instrument's scores represent the truth. On the other hand, there are various methods of arriving at estimates of reliability. Most of these methods involve comparisons of two scores for all of the individuals in some defined group. The relationship between the two sets of scores is expressed in terms of correlation coefficients. A correlation coefficient is a number that varies between -1 and +1 and that indicates the degree of relationship between variables. If the correlation coefficient is +1, there is perfect agreement between the two

variables. If one goes up, the other goes up a proportional amount; if one goes down, the other goes down a proportional amount. If the correlation coefficient is -1, the relationship is still perfect except reversed. If one variable goes up, the other goes down by a proportional amount. A correlation coefficient of 0 (zero) indicates no relationship between two variables.

There are several ways of estimating a test's reliability. Statistical formulas, such as Kuder-Richardson Formula 21, discussed in the reference articles, can be used when a single test is administered to a group. Reliability coefficients can also be arrived at by administering a single test to a group of people on two occasions and determining how well the scores correlate. Another way of estimating reliability is to follow the administration of one test with the administration of another test that has been constructed to be a parallel or equivalent version of the first test. The correlation between the two sets of scores is an estimate of reliability.

Often reliability estimates are obtained by comparing scores based on one-half of the questions in a test with scores based on the other half. This method is known as the split half reliability approach. The test is administered in ordinary fashion, but two scores are obtained on generally the odd-numbered and even-numbered questions. A correlation between the two halves gives the reliability of one-half of the test. To get the reliability of the whole test, the following formula is applied:

$$r_{xx} = \frac{2r_{nn}}{1 + r_{nn}}$$

where: r_{xx} = reliability of whole test

r_{nn} = reliability of 1/2 test

An important formula that is a generalization of the above will allow you to figure out the reliability of tests of increased length.

$$r_{kk} = \frac{k\, r_{xx}}{1 + (k-1)\, r_{xx}}$$

where: r_{xx} = reliability of a test of unit lenght

r_{kk} = reliability of a test made k times as long

An increase in the length of a test will result in higher reliability, if the questions that are added are similar to those already included. But it gets harder and harder to increase reliability by adding questions once you reach the upper levels of reliability.

The level of reliability needed for a test depends on the purpose for which that test will be used. A test used to make decisions about or give advice to individuals needs to have much higher reliability than a test that will be used only to characterize groups of people. Reversability of decisions is also important. If decisions based on test scores, such as grouping students for a particular unit of instruction, can be revised if they prove incorrect or harmful, less precision of measurement is necessary.

For general discussions of the concept of test reliability, consult basic textbooks for measurement courses such as:

Cronbach, L. J. Essentials of Psychological Testing (3rd ed.). New York: Harper and Row, 1970.

Thorndike, R. L., and Hagen, E. Measurement and Evaluation in Psychology and Education. (3rd ed.) New York: Wiley, 1969.

For an overview of theoretical and statistical approaches to reliability, see:

Stanley, J. C. Reliability In Robert L. Thorndike (Ed.) Educational Measurement (2nd ed.) Washington: American Council on Education, 1971.

Criterion-Referenced Tests

There have been a great number of articles produced in the past few years on the topic of criterion-referenced testing, many suggesting that these are new types of tests for which much of traditional measurement theory and practice is inappropriate. Before evaluating this position, it will be useful to review the meanings that have been assigned to the term "criterion-referenced tests."

Glaser and Nitko in a chapter in Educational Measurement offer the following definition of a criterion-referenced test:

> "A criterion-referenced test is one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards." (Glaser & Nitko, 1971, p. 653)

Glaser and Nitko go on to suggest that criterion-referenced tests can be differentiated from the more traditional norm-referenced tests in that they do not focus on the problem of individual differences and are not aimed at the task of determining an individual's relative standing in some norms group. Rather, they tell you what tasks an individual can or cannot do. Glaser and Nitko indicate that one step in the construction of a criterion-referenced test is the definition of a population of tasks. Some samples of populations of tasks are all possible pairs of two-digit numbers that might be added or a list of words all of which would have to be spelled.

Many of the articles on the subject of criterion-referenced tests have made use of the Glaser definition, but it is not the only one availa-

ble. Ebel (1971), for example, has characterized criterion-referenced measurement as follows:

> "The essential difference between norm-referenced and criterion-referenced measurements is in the quantitative scales used to express how much the individual can do. In norm-referenced measurement the scale is usually anchored in the middle, on some average level of performance for a particular group of individuals. The units on the scale are usually a function of the distribution of performance above and below the average level. In criterion-referenced measurement the scale is usually anchored at the extremities, a score at the top of the scale indicating complete or perfect mastery of some defined abilities, one at the bottom indicating complete absence of those abilities. The scale units consist of subdivision of these total score ranges." (Ebel, 1971, p. 282)

Both the Glaser and the Ebel statements contribute perspectives on the term "criterion-referenced." Their definitions contrast criterion-referenced and norm-referenced tests.

Still another view of criterion-referencing is provided by Popham and Husek (1969, p. 2):

> "Criterion-referenced measures are those which are used to ascertain an individual's status with respect to some criterion; i.e., performance standard. It is because the individual is compared with some established criterion, rather than other individuals, that these measures are described as criterion-referenced. The meaningfulness of an individual score is not dependent on comparison with other testees. We want to know what the individual can do, not how he stands in comparison with others."

It is interesting to note that these various definitions agree in that they emphasize the direct interpretability of scores on criterion-referenced tests, but differ in the extent to which they make reference to the method by which the test is constructed. Ebel emphasizes the scale from which interpretations are to be made. Other writers have taken the Glaser and Nitko position that the method of construction is central; Jackson (1970, p. 3), for example, states:

> "...the term 'criterion-referenced' will be used here to apply only to a test <u>designed</u> and <u>constructed</u> in a manner that defines explicit rules linking patterns of test performance to behavioral referents."

The definition of a criterion-referenced test as one that is specifically constructed to show what individuals can do leads to the development of tests by defining populations of tasks and then choosing representative samples from these populations. The narrower the definition of a population of tasks, the more homogeneous the tasks will be and the greater the degree of confidence you have about inferring how a student would perform on the total population of tasks, judging from his performance on a sample of these tasks. Because of the dependence of this method of criterion-referencing on the ability of the test constructor to specify a limited population of tasks, it seems most appropriate to situations wherein the number of tasks is limited by the nature of the subject matter -- e.g., identification of the letters of the alphabet -- or where the domain can be specified with reference to particular instructional materials -- e. g., the content of subunit ten of the text used by a particular class. Criterion-referencing by sampling from a fixed population seems most clearly appropriate to classroom developed tests or to special situations that have clearly defined limits.

Some writers have argued that only a sample of tasks directly associated with a particular learning objective can permit generalization to the objective. However tasks that are not actual samples may provide a good basis for generalization to an objective, once the basis for interpretation has been established. For example, performance on vocabulary and reading comprehension tests correlates very highly. Vocabulary test performance can be used, therefore, to make generalizations about students' reading comprehension skills. More generally, a sample of tasks covering a number of objectives can permit sound inferences to whole classes of objectives, including many not represented in the sample. This topic is treated in some

detail in a report entitled "Criterion-Referenced Interpretations of Survey Achievement Tests" (Fremer, 1972), available from Educational Testing Service, Princeton, New Jersey 08540.

Criterion Referenced Tests and "Other" Tests

The term "criterion-referenced" has been used almost interchangeably with "objectives-referenced" by some writers. An effort is often made to contrast these terms and the concepts underlying them with other testing terms that have a longer history of use. The practice of contrasting terms can lead to greater clarity of definitions, but it can also obscure significant relationships. There is much in common, for example, among the following terms and concepts:

criterion-referenced (or objectives-referenced) tests

diagnostic tests

mastery tests

minimum competency tests

What is a diagnostic test but a test that tells you where a student is strong or weak? If you can tell merely from a student's performance whether or not he is performing adequately in an area, then you must have some predetermined criteria for judging adequate or inadequate performance. Similarly, in the area of mastery or minimum competency tests, if you can define a performance on a series of tasks that you will accept as constituting mastery or minimum competency, then there is no need to compare one individual's score or rating with that of others. You need only look at his score to know whether or not he has reached mastery, or putting it another way, whether or not he has reached the criterion.

There are some differences between the ideas of criterion-referenced tests and diagnostic tests in that the clear implication behind the term diagnostic test is that someone is being tested so that an undesirable condition that is discovered can be treated and possibly corrected. No such plan of treatment is necessarily implied by the concept of a criterion-referenced test. A group of typists could be tested to see which of them can meet the criteria for successful office performance. One criterion for selecting secretaries might be a typing speed of 50 words per minute with no errors. If someone fails to meet that criterion, no additional training might be prescribed; he or she will just not be hired for the job.

In the classroom setting, most teacher-made tests involve some criterion-referenced and some norm-referenced interpretation. When a teacher selects the questions to ask on a test, he or she has in mind a level of performance that will be considered adequate for these questions. This constitutes a criterion for judging the achievement of students. On the other hand, the performance of the students will influence the criterion; it is not an absolute one. If everyone gets every question wrong on the test, the teacher is likely to see the need for readjusting his or her expectations. The students' performances serve as the basis for arriving at a criterion. Given a great deal of experience with many students and given past information on their performance on similar questions or even the same questions, the teacher may be much less willing to adjust a criterion to take into account low student performance in any one class. Rather, he or she may be willing to conclude that none of the students has met the criterion for adequate performance in a subject. Turning the issue around and placing the responsibility on the teacher as the source of instruction rather than on the students as learners, the teacher might conclude that he or she did not do as good a job this year

as last year because the students have not reached the specified level of performance.

An analysis of the steps required for development helps to reinforce the impression of similarities between criterion-referenced tests and other tests. Consider the questions -- Why am I testing? What should I test? Whom am I testing? What kinds of questions should I use? How long should my test be? How difficult should the test be? These same questions in some form will apply to any testing project. The particular purpose for testing will, of course, influence the answers to such questions. If, for example, the criterion of interest is performance on a particular population of tasks, then the question "What should I test?" might be answered in part by saying that it would be desireable to sample from that population of tasks. It would still be necessary in most situations, though, to decide how to test for mastery of tasks, so a heavy burden of decision would still rest with the test-maker. The issue of test difficulty will also be approached differently if you are only interested in whether or not students have achieved a particular criterion rather than how their performances compare. Even in this instance, however, the distinction between the criterion-referenced approach and the so-called traditional approach is easily exaggerated. Standards for people have to be determined by what people can do. We tend not to be interested in measuring what everyone can do or what no one can do because this information does not figure in many significant educational decisions. Instead our interest tends to focus on the large majority of settings where only some people know certain material or can do certain tasks. People do vary and it is this variation that influences all testing, whatever its label.

Criterion-Referenced Tests -- References

Ebel, Robert L. Criterion-referenced measurements: limitations. School Review, 1971, 79, 282-288.

Fremer, John Criterion-Referenced Interpretations of Survey Achievement Tests, TDM-72-1, Princeton, New Jersey: Educational Testing Service, 1972.

Glaser, Robert, & Nitko, Anthony J. Measurement in learning and instruction. In Robert L. Thorndike (Ed.), Educational Measurement, Washington, D. C.: American Council on Education, 1971. Pp. 625-670.

Jackson, Rex Developing criterion-referenced tests. TM Report No. 1. Princeton, New Jersey: ERIC Clearinghouse on Tests, Measurement, & Evaluation, 1970. (ED041 052; 18p.)

Popham, James W., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

APPENDIX C - "Supplementary Materials - March 31, 1972 version",
Final Report - "Preparation of a Filmstrip Unit on Basic Measurement Principles"
October 31, 1973, Project No. 0-9050, Contract No. OEC-0-70-4777

Supplementary Materials for Filmstrip -- "Planning a Test"[1]

by

John Fremer[2]

Educational Testing Service

The topics covered in this document are:

- Learning How to Develop Tests
- Obtaining Information About Tests
- Preparing a Test Plan
- Kinds of Test Questions: Advantages and Disadvantages
- Reliability
- Criterion-Referenced Tests

Learning How to Develop Tests

The filmstrip "Planning a Test" provides an overview of the test development process and calls attention to a number of significant issues for the test planner and developer. If you have already taken a course in Tests and Measurements, the filmstrip provides a review of material that is familiar to you. If you are only now taking such a course or if you have no training in measurement, the filmstrip should make you aware that there is much to be learned about test development. If it is successful, the filmstrip may whet your appetite for further investigation of the topics that were covered. For those who want to increase their understanding through independent reading, a highly recommended first step is to obtain a good elementary measurement textbook. There are a number available including:

---

[1]The filmstrip "Planning a Test" and these supplementary materials were prepared by Educational Testing Service under Contract No. OEC-0-70-4777 (Project No. 0-9050) with the U.S. Department of Health, Education, and Welfare, Office of Education, National Center for Educational Research and Development. Single copies of supplementary materials in a form suitable for photographic reproduction will accompany each filmstrip order. [Address for orders: Educational Testing Service, Princeton, New Jersey 08540.] These supplementary materials may be reproduced without permission from the Office of Education or Educational Testing Service.

[2]Substantial contributions to the development of these supplementary materials were made by: Clair Bowman, Miriam Bryan, Eleanor Horne, S. Donald Melville, and Michael Ziky.

Cronbach, L. J. Essentials of psychological testing. (3rd ed.) New York: Harper & Row, 1970.

Thorndike, R. L., and Hagen, E. Measurement and evaluation in psychology and education. (3rd ed.) New York: Wiley, 1969.

If you want a more comprehensive treatment of major measurement topics than can be given in an elementary textbook, the best single source of authoritative articles and references that can serve as a guide to further study is:

Thorndike, R. L. (Ed.) Educational Measurement. Washington: American Council on Education, 1971.

In addition to an introductory chapter by the editor Robert L. Thorndike, the major areas covered by this book are:

Test Design, Construction, Administration, and Processing (7 chapters)

Special Types of Tests (3 chapters)

Measurement Theory (5 chapters)

Application of Tests to Educational Problems (4 chapters)

For a more complete listing of available textbooks and of other source and reference materials in the area of educational measurement, you can write Educational Testing Service, Princeton, New Jersey, 08540 for the following pamphlet:

Locating Information on Educational Measurement: Sources and References. (2nd ed.) Princeton: Educational Testing Service, 1969.

It is also possible to obtain from The Psychological Corporation, 304 East 45th Street, New York, New York 10017, a listing of a series of documents on testing issues entitled the Test Service Bulletins and from Harcourt, Brace Jovanovich, Inc., 757 Third Avenue, New York, New York 10017, a list of Test Service Notebooks that discuss topics related to test theory, administration of testing programs, proper use of test results, and results of research studies.

Obtaining Information About Tests

There is a great deal of information available about tests and other measurement devices, but much of it is not readily accessible to the individual teacher or researcher who is about to embark on a particular test development project. Since constraints such as time and availability

of materials will vary from individual to individual, it is not possible to identify the ideal search pattern for all situations. An attempt will be made, however, to identify a comprehensive approach to the problem of searching for background information. If several of the recommended major sources are used, it should be possible to collect a great deal of information in those areas where many instruments have been developed.

The following discussion will focus on four categories of suggested procedures for a systematic search for tests and assessment devices.

I. Preparatory Activities
II. Survey of Test-Centered Literature
III. Literature Search
IV. Contacts in the Field

I. Preparatory Activities

No search is likely to be very productive unless you spend a good deal of time first deciding what it is that you are looking for. Some important steps to consider when you are contemplating a search for information about tests or other measurement instruments are:

A. What are your testing objectives and how do they relate to the overall objectives of your project?

What type of information are you seeking and how will you use it?

What decisions do you have to make and what part can test-based information contribute to these decisions?

B. What are the constraints that will influence your search and subsequent test development?

1. Time and Schedule -- When do you need to have your test or questionnaire? How much time can you devote to your search and to the task of development?
2. Available Personnel -- Are you "on your own" or do you have other people working with you? If you have help, you should consider ways of dividing responsibility.
3. Fiscal Considerations -- Can you afford to purchase materials about tests or to design a testing approach that requires expensive equipment?

4. Nature of the Population -- Whom will you be testing? Consider such factors as age, previous familiarity with tests, language development, and motivation to cooperate.
5. Available Facilities -- What search sources are convenient such as a nearby college library? What sources can be reached only with considerable effort?

C. Is there a current comprehensive summary or state-of-the-art publication in your area of interest?

Your own search will be greatly assisted if you can locate a recent book or article that summarizes much of the recent work in your area of interest. You can identify for further study the references that relate most closely to your interests, and you can usually determine what publications would be likely to publish additional articles relevant to your project.

D. Design a worksheet or checklist that you can use in recording information which will be useful in the comparison or selection of measures.

II. Survey of Test-Centered Literature

The task of reviewing a field is best handled by referring to one or more of the summaries that are available. You can learn what tests exist and read critical reviews of these tests by individuals in the field.

A. Mental Measurements Yearbook Series (Gryphon Press, Highland Park, New Jersey)

1. Mental Measurements Yearbooks (MMY)
2. Tests in Print
3. Reading Tests and Reviews
4. Personality Tests and Reviews

This series includes descriptions of tests, critical reviews, publishers' directories, and bibliographical references.

B. CSE: Elementary School Test Evaluations & CSE--ECRS Preschool/Kindergarten Test Evaluations

These volumes include ratings of tests on a number of criteria. They are published by the Center for the Study of Evaluation and the Early Childhood Research Center, UCLA Graduate School of Education, Los Angeles, California.

C. NCME Measurement News

This newsletter of the National Council on Measurement in Education contains general articles on testing issues as well as announcements of new tests and lists of test reviews.

D. Test Collection Bulletin (TCB) -- ETS, Princeton, New Jersey

This is a quarterly digest of information on tests and services which generally have become available after the publication of the most recent Mental Measurement Yearbook. Commercially available and "research" measures are described. The Bulletin does not evaluate the tests that are listed, but it does provide references to test reviews.

E. Promotional Materials from Test Publishers

Check publisher's catalogs and announcements for references to tests, services, and technical data on specific measures made available after the publication of the Mental Measurements Yearbook.

F. Other Appropriate Reference Materials

A document entitled "An Annotated Bibliography of References to Tests and Assessment Devices" may be obtained without charge from the ETS Test Collection, Educational Testing Service, Princeton, New Jersey, 08540.

III. Literature Search

If you want to locate the most recent articles on the development and use of tests in your area of interest, some use of reference sources that abstract or index articles, project reports, etc., is in order. In addition, it is often useful to search the most recent issues of those journals that tend to have many articles listed in the summaries of past work.

A. Traditional Reference Tools

These sources of abstracts of educational and psychological literature can usually be found on the reference shelves of college libraries.

1. Psychological Abstracts
2. Education Index
3. Research Studies in Education
4. Dissertation Abstracts

B. Educational Resources Information Center (ERIC)

ERIC is actually a series of clearinghouses, each specializing in a specific area of education. Reports and articles are collected and indexed in ERIC publications: 1. Research in Education, 2. Current Index to Journals in Education, and 3. Clearinghouse Publications (frequently a source of state-of-the-art papers). Items cataloged in the ERIC system can be ordered in microfiche or hard copy. The ERIC Clearinghouse on Tests, Measurements & Evaluation is located at Educational Testing Service in Princeton, New Jersey.

C. Professional Journals

The journals in an area not only provide direct reports of developmental and research activities, but also related theoretical papers. They often provide test reviews which may also be referenced in the Test Collection Bulletin and the NCME Measurement News. The advertisements in the journals are an excellent source of information on new materials and services available from commercial publishers and research organizations.

D. Texts and References in the Area Under Study

Are tests included, described, or mentioned? Scan the bibliographies for additional sources of information.

IV. Contacts in the Field

A. Use of Information Directories

Particularly useful is the Encyclopedia of Information Systems and Services, edited by Anthony Kruzas. (New York: Edward Brothers, Inc. c 1971)

B. Educational Testing Service Test Collection

C. Head Start Test Collection (ETS, Princeton, New Jersey)

(Funded by the Office of Child Development, Department of Health, Education, and Welfare)

Both of these test collections provide on-site, telephone, and mail reference services. You can write for lists of publications that summarize available instruments in specified areas.

D. Research and Development Centers and Child Study Laboratories, etc.

E. Professional Organizations and Special Interest Groups such as:

American Printing House for the Blind,

International Reading Association,

American Association for Health, Physical Education and Recreation

Preparing a Test Plan

Even before searching for information about tests it is wise to give careful thought to your objectives for testing and to develop a preliminary test plan. After you have exhausted the sources of information about tests that are available to you, however, and have resolved to undertake the development of your own tests, there remains a need for a good deal of careful and structured planning. An extremely useful source of guidance in this connection can be the chapter entitled, "Planning the Objective Test" by Sherman Tinkelman in Educational Measurement, edited by R. L. Thorndike. Other sources of information on preparing tests include:

Ebel, R. L. Measuring Educational Achievement. Englewood Cliffs, New Jersey: Prentice-Hall, 1965.

This text deals primarily with achievement tests prepared by teachers and professors for use in their own classes. It emphasizes methods of test development and item analysis, rather than the selection and use of standardized tests. Included are practical suggestions for planning, constructing, administering, and scoring classroom tests and for analyzing the results. No previous special training in educational measurement is assumed.

Educational Testing Service Making the Classroom Test: A Guide for Teachers. (2nd ed.), Princeton: Educational Testing Service, 1961.

This pamphlet reviews the plans and procedures used by four hypothetical teachers to prepare good tests and considers a number of special problems faced in the writing and scoring of tests.

If you are primarily concerned with the measurement of attitudes and interests rather than with measuring knowledge, the following books can be useful:

Edwards, Allen L. Techniques of Attitude Scale Construction. New York: Appleton-Century-Crofts, 1957.

The author notes that his book, "...is intended for those who may desire to measure attitudes toward something in which they are interested, but who fail to find an appropriate scale available."

Robinson, John P., and Shaver, Phillip R. Measures of Social Psychological Attitudes. Ann Arbor, Michigan: Institute for Social Research, The University of Michigan, 1969.

A review and evaluation of 112 empirical scales for measuring social psychological attitudes such as: life satisfaction and happiness, self-esteem, dogmatism, sociopolitical attitude, social values, attitudes toward people, and religious attitudes. A copy of each scale is included.

Shaw, M. E., and Wright, J. M. Scales for the Measurement of Attitudes New York: McGraw-Hill, 1967. 604 pp.

Not addressed to a particular audience, this book brings together a number of useful scales suitable for research purposes and group testing. The authors caution against the use of these scales for individual measurement, diagnosis, or personnel selection. Topics include the nature of attitudes and methods of scale construction; the scales, presented in eight chapters; and evaluation and suggestions for improvement.

One of the major components of test planning is the identification of appropriate content for the test. Since the validity of a test is dependent upon the extent to which it actually measures what it is supposed to measure, you need to obtain a representative and balanced sample of tasks or questions from the universe of content to be covered. When developing achievement tests in subject-matter areas, it will be useful to pose questions such as the following:

What are the important things you would expect a person who has studied this subject to know?

What intellectual skills should he have acquired?

What level of understanding of the material should he be required to demonstrate?

What is the relative importance of these various elements?

Whenever possible, you should supplement your subjective impressions regarding content with factual information gathered from an analysis of the materials that were used in instruction in the particular subject-matter area. Among the sources of information that might be used are:

Textbooks and teacher's manuals -- When preparing a test for your own course, a review of this material will be greatly facilitated by your knowledge of just what assignments were particularly stressed. When preparing tests for courses that you have not yourself taught, as would be the case in a departmental testing or research setting, it will be necessary to not only review assignment sheets but to determine precisely what sections of the material were actually covered.

Course syllabi and curriculum guides -- Again for most purposes, you need to differentiate between what should have been covered and what actually was covered. Only in the unusual situation where you are addressing the evaluation of a course as a type of outside auditor who can only "go by the book" should you proceed as though a pre-course plan has actually been followed.

Lesson plans, lecture notes, laboratory activities, films, and filmstrips -- See preceding comments.

The content categories that you develop should be comprehensive ones covering all areas of interest. It will be useful to group the material into meaningful clusters and to determine the emphasis to be given each cluster in the final test.

In addition to identifying the content to be covered by a test, you will also need to address the issue of the skills or level of understanding that you will require of students. Each of the content areas will contain factual material as well as material that could serve as the basis for questions requiring inference and the analysis of relationships. When considering the

different skills or abilities to be tested, it will be useful to start by reviewing the appropriate one of the following taxonomies of educational objectives:

1. Bloom, B. S. (Ed.) Taxonomy of Educational Objectives: The Classification of Educational Goals; Handbook I: Cognitive Domain. New York: David McKay Co., Inc., 1956. 207 pp.

This volume is intended for educators and research workers who deal with curriculum and evaluation problems. It provides a classification of the cognitive goals in education; i.e., those goals which primarily involve intellectual considerations. Part I explains the nature and development of the taxonomy and describes the principles and problems in classifying educational objectives; Part II presents the hierarchical taxonomy and illustrative materials for each level: knowledge, comprehension, application, analysis, synthesis, and evaluation.

2. Krathwohl, D. R., Bloom, B. S., and Masia, B. B. Taxonomy of Educational Objectives: The Classification of Educational Goals; Handbook II: Affective Domain.

This second book in the Taxonomy series is devoted to the affective goals of education; i.e., those goals which primarily concern emotional or feeling behaviors of students such as appreciation, attitudes, and values. Part I describes the nature of the affective domain and the classification structure and describes the evaluation of affective objectives at each level of the structure. It also analyzes the relation of the affective to the cognitive domain.

The Cognitive Domain

While either the cognitive or affective or a third domain, the psychomotor, may be represented among the objectives for teaching or the variables in research, a review of the cognitive domain permits the explication of principles that apply to any of the areas. It was for this reason that the filmstrip used the cognitive skills dimension as one of the two dimensions of the content-skills matrix.

As noted in the filmstrip, the cognitive domain is subdivided into six categories: knowledge, comprehension, application, analysis, synthesis, and

evaluation. A further explanation of each is in order.

The knowledge level may be equated with recall of information, ranging from recall of specific bits of information to the recall of universals and abstractions in a field. Comprehension incorporates most of what teachers have often intended by their use of the word understand -- a level of understanding such that the individual knows what is being communicated by a stimulus but is not necessarily aware either of the implications of the stimulus or how it relates to other material. Application involves making use of an abstraction learned in one setting in other, different situations.

An individual is operating at the analysis level when he breaks down a stimulus into component parts in order to indicate how it is organized, what effects it manages to convey, how it conveys them, etc. Synthesis is the reverse of this process. Here the individual takes elements and puts them together in such a manner as to form a unified whole which is something other than a pattern or object which existed prior to his efforts. Evaluation, as the word clearly implies, is making judgments of value about the extent to which stimuli satisfy criteria.

These six categories appear to be logically hierarchical in nature, with evaluation requiring all of the other skills, synthesis requiring all but evaluation, and so forth. Empirical evidence with which to either verify or reject the logical evidence is largely lacking. This does not detract greatly from the utility which the classification system has to offer; it is merely a caution of which you should be aware. One further caveat is in order. Note that for any given individual, the category system is designed only to indicate what he is intended to be doing at a given moment. What the teacher or researcher intends and what the individual is actually doing may be quite different things. There are obvious examples. When an individual has already had an experience in which he has had to evaluate the worth of a particular argument, to have him re-evaluate the same argument against the same criteria a second time is merely an exercise in recall of information -- knowledge level behavior. He must be using a different criterion or evaluating a different argument for his behavior to actually be evaluative in nature.

The two volumes published to date provide full descriptions of levels within the cognitive and affective domains and provide for your assistance examples of both the type of objectives characteristic of each level and the nature of test questions which assess subjects' competence in them. Both of these volumes contain a summary form of the levels within their respective domains which you will find most useful as you attempt to categorize objectives or to write test questions.

Kinds of Test Questions: Advantages and Disadvantages

A number of considerations will influence the selection of questions for a test. Obviously, you will want to use a format that gets at the kinds of knowledge or skills that you are particularly interested in. To some extent, the nature of the subject-matter will influence your choices as can be illustrated with the following examples:

1. If you want to know how a student feels about some issue, you will often find it efficient to ask his degree of agreement with statements relating to that issue.
2. If you are interested in the student's ability to translate from another language into English, your questions will have to include material in that language.
3. If you are interested in evaluating a student's map-reading skills, the use of maps in your questions is indicated.

In some situations you may want to consider the use of essay questions, perhaps because they require a student to develop answers from his own background, without the benefit of suggested possibilities, and to express the answer in his own words. You may also be attracted by the fact that essay tests can be easily prepared. There are fewer questions to write and, if necessary, they can be written on a classroom chalkboard. Moreover, the use of essay questions very largely eliminates guessing.

It is true that essay questions <u>can</u> serve to measure some higher level abilities when pupils are required to present evidence, evaluate, analyze, solve new problems, or approach problems in a new way.

Unfortunately, too many so-called essay questions do not do that. These are of the "Name the six largest cities in the United States" or "List the characters in each of the following stories" variety, types of questions which would be less than mediocre if presented in a choice response format. Or essay questions may be ambiguously stated or be stated so generally that a pupil can bluff or "talk around" the subject. Poor essay questions of the type described are easy to write. Writing more challenging essays requires considerable thought.

The principal disadvantage of the essay question is, of course, the unreliability of the scoring of the answer. Why is it so difficult to achieve reliability in scoring the answers to essay questions? One reason is that judgments differ. One grader may think that an answer is good; another may think that it is poor. This may not be the result of the inadequacy of one of the graders; it may be the result of an honest difference of opinion on the relative merits of an answer. For another thing, if the essay question is concerned with a controversial topic, the grader's judgment is likely to be influenced by his own convictions. In the classroom setting, some teachers believe that the pupil's ability to express himself must be taken into account; these teachers deduct credit on answers to essay questions in social studies, science, and other subject-matter areas for poor English expression. Other teachers believe that a pupil should not be penalized if he knows the subject-matter but cannot express himself especially well. Then, too, the teacher's judgment may be influenced by how a paper looks -- the easier the paper is to read, the higher is the score assigned. Or the teacher's judgment may be influenced by the "halo" effect. There may be a tendency to mark in terms of work in class or even in terms of what is expected of the class to which the pupil is assigned. The "halo" effect may also operate from question to question, the quality of the answer to the first question influencing the scoring of the answers to subsequent questions. Add to all of these sources of unreliability of scoring the fact that teachers frequently correct tests after a long school day, even late in the evening, and that their scoring

is likely to be somewhat erratic as a result of real fatigue. Sometimes teachers who rescore essay questions after an interval of time find themselves coming up with quite different scores.

Beyond the problem of reliable scoring, there are two other limitations to essay questions which should be mentioned. First, there is the possibility of inadequacy of sampling when only a few questions are selected to cover a large content area. Then, there is the large penalty per question that will result if the pupil does not know the answer to the question or, even more serious, if he knows the answer but does not understand the question.

Finally, a choice of essay questions is frequently permitted. If the performance of pupils is to be compared one with another, a choice among essay questions should not be permitted. Without elaborate equating of the questions on the basis of their difficulty to the pupils who attempt them, there is no way of knowing, for example, how well the pupils who chose to answer questions 1, 2, and 4 would have done had they chosen to answer questions 3, 5, and 6 instead. And the better pupils who attempt the more challenging questions sometimes write less acceptable answers than do the less talented pupils who are satisfied to answer the easier questions.

To sum up the discussion of essays, although they are appropriate in some circumstances, they tend to be an extremely inefficient method of obtaining information about individuals. They are obviously useful as measurement devices only when multiple, independent grading is arranged, using graders who can agree in advance on common criteria. For a comprehensive discussion of essay examinations, see the chapter by Coffman in Educational Measurement.

Now what about questions of the objective type? They have at least four advantages:

1. They permit wider sampling of learning in a relatively short time, making spotty preparation on the part of the pupil more obvious.
2. They can be reliably scored. If the items are unambiguous and the test has been keyed properly, the scoring errors will be clerical errors rather than errors of judgment.

3. They are more easily scored. Scoring time is reduced. The teacher is freed from suspicion of partiality. Frequently, they can be scored by the pupils themselves.
4. They lend themselves rather readily to item analysis. The teacher can, over a period of time, assemble a file of questions, retaining from each test the questions that are of proper difficulty, and that discriminate well between high and low achievers.

One limitation of questions of the objective type is that they are difficult to construct if they are to test anything more than memory. All types of objective questions, especially multiple-choice questions, can test for recognition of assumptions, interpretation of data, recognition of limitations, application of principles, and a variety of other higher intellectual abilities and skills. Experience with several types of objective questions, however, suggests that some are more efficient for such purposes than others. Because the construction of objective questions that do measure understanding and thinking requires a large amount of time and considerable ingenuity, teachers are likely to be content with questions testing principally knowledge of facts -- sometimes very trivial facts. This limitation of questions of the objective type, then, is not a limitation inherent in the questions themselves but a limitation imposed by the test constructor.

Five types of objective questions are popular with classroom teachers -- completion, true-false, matching, classification, and multiple-choice. All of these types can serve useful purposes if the questions are well drawn and the answers to be supplied or chosen are definite and specific. They are all easily adapted to classroom situations and a large number of them can be administered in a relatively short time. In most testing situations, a balanced blend of several types of objective questions will prove most effective. When more than one type is used, however, it is a good idea to group together questions of the same type so that the pupils will not be confronted with too frequent changes in directions.

The textbooks already mentioned (Cronbach, Ebel, and Thorndike and Hagen) provide guidance on types of questions and how to write them. In

addition, you may want to look at the following:

Gerberich, J. R. Specimen Objective Test Items: A Guide to Achievement Test Construction. New York: Longmans, Green and Co., 1956.

This book is designed primarily for teachers. It serves the test specialist by providing several systematic classifications of instruments and techniques used in achievement testing. The coverage of the book is as follows: measurement of educational achievement (1 chapter); specimen objective achievement test items (11 chapters); classification of objective achievement test items (3 chapters); and tests, nontest tools, and techniques used in achievement measurement (2 chapters).

Sanders, N. M. Classroom Questions: What Kinds? New York: Harper and Row, 1966.

Intended for teachers, particularly social studies teachers, and other makers of classroom test questions, this book is based on Bloom's Taxonomy of Educational Objectives. Chapter titles include: "Questions Designed for More Than Memory"; "Memory"; "Translation"; "Interpretation"; "Application"; "Analysis"; "Synthesis"; "Evaluation"; and "Planning for Questioning."

Reliability

Reliability is a term for the dependability of a measurement. If we could measure the same set of people again and again with the same or comparable instruments or procedures, would we get the same or similar results on each testing? The answer is "no". Measurement, like any other human endeavor, involves a certain amount of error. The errors can be of two types: there is systematic error, such as when scores on an instrument are all biased in one direction -- for example, a ruler that is too short, a scale that adds five pounds to each weighing, a test that has a miskeyed question. Scores on such defective instruments may be wrong, but they will be wrong consistently, dependably, predictably, and reliably. Random errors, however, as the name implies, occur by chance. They are temporary and shifting, often due to unknown causes such as fluctuations in mood, health, motivation, and so on. The greater the amount of random error in a

measuring instrument, the less the reliability of that instrument.

To help us think about reliability, a good example would be that of two watches -- one very fine and expensive, the other one rather cheap and poorly made. The fine watch will measure time reliably and dependably. Even if it is set to the wrong time so that it does not give an accurate estimate of the correct time, it will still be consistent, or, in this sense, "reliable." The cheap watch, on the other hand, will have a certain amount of purely random error. Sometimes it will be too fast, sometimes it will be too slow, and it may be impossible to predict which way the error is directed. It is this kind of random, unpredictable error that makes the cheap watch unreliable.

One very good definition of reliability is that reliability is a proportion. It is the proportion of the score that is a true measurement of the thing being measured. Obviously, the more random error, the smaller the proportion of truth in a measurement. The less truth in the measurement -- the smaller the proportion of truth to error -- the lower reliability will be.

Unfortunately, there is no direct way of determining reliability by discovering what proportion of a measuring instrument's scores represent the truth. On the other hand, there are various methods of arriving at estimates of reliability. Most of these methods involve comparisons of two scores for all of the individuals in some defined group. The relationship between the two sets of scores is expressed in terms of correlation coefficients. A correlation coefficient is a number that varies between -1 and +1 and that indicates the degree of relationship between variables. If the correlation coefficient is +1, there is perfect agreement between the two variables. If one goes up, the other goes up a proportional amount; if one goes down, the other goes down a proportional amount. If the correlation coefficient is -1, the relationship is still perfect except reversed. If one variable goes up, the other goes down by a proportional amount, A correlation coefficient of 0 (zero) indicates no relationship between two variables.

Reliability estimates are most often obtained by applying statistical formulas such as Kuder-Richardson Formula 21, discussed in the reference articles, to the score distributions obtained by administering a single test to a specified population. Reliability coefficients can also be arrived at by administering to a group of people a single test on two occasions or one test followed after some interval by a parallel form of the test. The time interval chosen depends on what the test is to be used for. If one wants to make long-range decisions based on the test, it seems reasonable to allow a relatively long time period between testing and retesting in order to get a reasonable estimate of the stability of the scores obtained. If, on the other hand, one is interested in making very short-range predictions, it seems reasonable to allow only a short time between testing and retesting.

When the correlation between two forms is used as an estimate of reliability, the degree of similarity between the two forms determines the correlation. Often estimates are obtained by comparing scores based on one-half of the questions in a test with scores based on the other half. This method is known as the split half reliability approach. The test is administered in ordinary fashion, but two scores are obtained on generally the odd-numbered and even-numbered questions. A correlation between the two halves gives the reliability of one-half of the test. To get the reliability of the whole test, the following formula is applied:

$$r_{xx} = \frac{2r_{nn}}{1 + r_{nn}}$$

where: r_{xx} = reliability of whole test

r_{nn} = reliability of 1/2 test

An important formula that is a generalization of the above will allow you to figure out the reliability of tests of increased length.

$$r_{kk} = \frac{k\, r_{xx}}{1 + (k-1)\, r_{xx}}$$

where: r_{xx} = reliability of a test of unit length

r_{kk} = reliability of a test made k times as long

This formula makes it clear that an increase in the length of a test will result in higher reliability, if the questions that are added are similar to those already included. But it gets harder and harder to increase reliability by adding questions once you reach the upper levels of reliability.

The level of reliability needed for a test depends on the purpose for which that test will be used. A test used to make decisions about or give advice to individuals needs to have much higher reliability than a test that will be used only to characterize groups of people. Reversability of decisions is also important. If decisions based on test scores, such as grouping students for a particular unit of instruction, can be revised if they prove incorrect or harmful, less precision of measurement is necessary.

For general discussions of the concept of test reliability, consult basic textbooks for measurement courses such as:

Cronbach, L. J. Essentials of Psychological Testing (3rd ed.). New York: Harper and Row, 1970.

Thorndike, R. L., and Hagen, E. Measurement and Evaluation in Psychology and Education. (3rd ed.) New York: Wiley, 1969.

For an overview of theoretical and statistical approaches to reliability, see:

Stanley, J. C. Reliability In Robert L. Thorndike (Ed.) Educational Measurement (2nd ed.) Washington: American Council on Education, 1971.

Criterion-Referenced Tests

There have been a great number of articles produced in the past few years on the topic of criterion-referenced testing, many suggesting that these are new types of tests for which much of traditional measurement theory and practice is inappropriate. Before evaluating this position, it will be useful to review the meanings that have been assigned to the term "criterion-referenced tests."

Glaser and Nitko in a chapter in Educational Measurement offer the following definition of a criterion-referenced test:

> "A criterion-referenced test is one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards." (Glaser & Nitko, 1971, p. 653)

Glaser goes on to suggest that criterion-referenced tests can be differentiated from norm-referenced tests in that they do not focus on the problem of individual differences and are not aimed at the task of determining an individual's relative standing in some norms group. Rather, they tell you what an individual can or cannot do. Glaser talks about the need to construct a criterion-referenced test by defining a population of tasks. Some samples of populations of tasks are all possible pairs of two-digit numbers that might be added or a list of words all of which would have to be spelled.

Many of the articles on the subject of criterion-referenced tests have made use of the Glaser definition, but it is not the only one available. Ebel (1971), for example, has characterized criterion-referenced measurement as follows:

> "The essential difference between norm-referenced and criterion-referenced measurements is in the quantitative scales used to express how much the individual can do. In norm-referenced measurement the scale is usually anchored in the middle, on some average level of performance for a particular group of individuals. The units on the scale are usually a function of the distribution of performance above and below the average level. In criterion-referenced measurement the scale is usually anchored at the extremities, a score at the top of the scale indicating complete or perfect mastery of some defined abilities, one at the bottom indicating complete absence of those abilities. The scale units consist of subdivisions of these total score ranges. (Ebel, 1971, p. 282)

Both the Glaser and the Ebel statements contribute perspectives on the term "criterion-referenced." Their definitions contrast criterion-referenced and norm-referenced tests.

Still another view of criterion-referencing is provided by Popham and Husek (1969, p. 2):

> "Criterion-referenced measures are those which are used to ascertain an individual's status with respect to some criterion; i.e., performance standard. It is because the individual is compared with some established criterion, rather than other individuals, that these measures are described as criterion-referenced. The meaningfulness of an individual score is not dependent on comparison with other testees. We want to know what the individual can do, not how he stands in comparison with others."

It is interesting to note that these various definitions agree in that they emphasize the direct interpretability of scores on criterion-referenced tests, but differ in the extent to which they make reference to the method by which the test is constructed. Ebel emphasizes the scale from which interpretations are to be made. Other writers have taken the Glaser position that the method of construction is central; Jackson (1970, p. 3), for example, states:

> "...the term 'criterion-referenced' will be used here to apply only to a test designed and constructed in a manner that defines explicit rules linking patterns of test performance to behavioral referents."

The definition of a criterion-referenced test as one that yields direct criterion-referenced interpretations by virtue of the method by which it was constructed leads to the development of tests by defining populations of tasks and then choosing representative samples from these populations. The narrower the definition of a population of tasks, the more homogeneous the population will be and the greater the degree of confidence one will be able to have about an inference from performance on a sample of such tasks to the total population of tasks. Because of the dependence of this method of criterion-referencing on the ability of the test constructor to specify a limited population of tasks, it seems most appropriate to situations wherein the number of tasks is delimited by the nature of the subject matter -- e.g., identification of the letters of the alphabet -- or where the domain can be specified with reference to particular instructional materials -- e.g., the content of subunit ten of the text used by a particular class. Criterion-referencing by sampling from a fixed population seems most clearly appropriate to classroom developed tests or to special situations that have clearly defined limits.

Criterion-Referencing Through Validation for Specific Criteria

Direct inferences about what a test-taker can or cannot do -- criterion-referenced inferences, that is -- need not be restricted to tests that are composed of actual samples of the behaviors of interest. Considerable use can be made of the very high relationships that have been observed among many apparently diverse tasks within such global areas as reading, language usage, or mathematics. Although some writers have argued

that only a sample of tasks directly associated with a particular objective can permit generalization to that objective, other tasks that are not samples of that objective may provide just as good a basis for such a generalization, once the basis for interpretation has been established. More generally, a sample of tasks covering a number of objectives can permit sound inferences to whole classes of objectives, including many not represented in the sample. This topic is treated in some detail in a report entitled "Criterion-Referenced Interpretations of Survey Achievement Tests" (Fremer, 1972), available from Educational Testing Service, Princeton, New Jersey 08540.

Criterion-Referenced Tests and "Other" Tests

Although the definitions of criterion-referenced tests tend to emphasize differences between such tests and traditional tests, and sometimes to suggest that they constitute new types of tests, there are clear relationships among the meanings assigned to the term criterion-referencing and those assigned to other terms that have a longer history of use. There is much in common, for example, among the following terms and concepts:

criterion-referenced tests
diagnostic tests
mastery tests
minimum competency tests

What is a diagnostic test but a test that tells you where a student is strong or weak? If you can tell merely from a student's performance whether or not he is performing adequately in an area, then you must have some predetermined notion of what constitutes adequate or inadequate performance. Similarly, in the area of mastery or minimum competency tests, if you can define a performance on a series of tasks that you will accept as constituting mastery or minimum competency, then there is no need to compare one individual's score or rating with that of others. You need only look at his score to know whether or not he has reached mastery, or putting it another way, whether or not he has reached the criterion.

There are some differences between the ideas of criterion-referenced tests and diagnostic tests in that the clear implication behind the term diagnostic test is that someone is being tested so that any undesirable condition that is discovered can be treated and possibly corrected.

No such plan of treatment is necessarily implied by the concept of a criterion-referenced test. A group of typists could be tested to see which of them can meet the criteria for successful office performance. One criterion for selecting secretaries might be a typing speed of 50 words per minute with no errors. If someone fails to meet that criterion, no additional work might be prescribed; he or she will just not be hired for the job.

In the classroom setting, most teacher-made tests have some criterion-referenced and some norm-referenced interpretation behind them. When a teacher selects the questions to ask on a test, he or she has in mind a level of performance that will be considered adequate for these questions. This constitutes a criterion to be achieved by the students. On the other hand, the performance of the students will influence the criterion; it is not an absolute one. If everyone gets every question wrong on the test, the teacher is likely to see the need for readjusting his or her expectations. The student's performance serves as the basis for arriving at a criterion. Given a great deal of experience with many students and given past information on their performance on similar questions or even the same questions, the teacher may be much less willing to adjust a criterion to take into account low student performance in any one class. Rather, he or she may be willing to conclude that none of the students has met the criterion for adequate performance in a subject. Turning the issue around and placing the responsibility on the teacher as the source of instruction rather than on the students as learners, the teacher might conclude that he or she did not do as good a job this year as last year because the students have not reached the specified level of performance.

An analysis of the steps required for development helps to reinforce the impression of similarities between criterion-referenced tests and other tests. Consider the questions -- Why am I testing? What should I test? Whom am I testing? What kinds of questions should I use? How long should my test be? How difficult should the test be? These same questions in some form will apply to any testing venture. The particular purpose for testing will, of course, influence the answers to such questions. If,

for example, the criterion of interest is performance on a particular population of tasks, then the question "What should I test?" might be answered in part by saying that it would be desirable to sample from that population of tasks. It would still be necessary in most situations, though, to decide how to test for mastery of tasks, so a heavy burden of decision would still rest with the test-maker. The issue of test difficulty will also be approached differently if you are only interested in whether or not students have achieved a particular criterion or if you want to know only what proportion of a given set of content or skills they have learned. Even in this instance, the distinction between the criterion-referenced approach and the so-called traditional approach is easily exaggerated. Standards for people have to be determined by what people can do. We tend not to be interested in measuring what everyone can do or what no one can do because this information does not figure in many significant educational decisions. Instead our interest tends to focus on the large majority of settings where only some people know certain material or can do certain tasks. People do vary and it is this variation that influences all testing, whatever its label.

Criterion-Referenced Tests -- References

Ebel, Robert L. Criterion-referenced measurements: limitations. School Review, 1971, 79, 282-288.

Fremer, John Criterion-Referenced Interpretations of Survey Achievement Tests, TDM-72-1, Princeton, New Jersey: Educational Testing Service, 1972.

Glaser, Robert, & Nitko, Anthony J. Measurement in learning and instruction. In Robert L. Thorndike (Ed.), Educational Measurement, Washington, D. C.: American Council on Education, 1971. Pp. 625-670.

Jackson, Rex Developing criterion-referenced tests. TM Report No. 1. Princeton, New Jersey: ERIC Clearinghouse on Tests, Measurement, & Evaluation, 1970. (ED041 052; 18 p.)

Popham, James W., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Planning a Test

(Formerly, Specifications for Measuring Instruments)

1. Title[1]

2. What is a measuring instrument?

3. Tests and questionnaires are common examples in the educational world. As measuring instruments, they can provide valuable information, but only if they are developed and used properly.

4. Specifications are to a test, or to other measuring instruments, what blueprints are to a house: an essential first step upon which the quality and usefulness of the final product depends.

5. To develop your specifications, you need to consider many issues, among them the following six questions: why am I testing, what should I test, whom am I testing, what kinds of questions should I use, how long should my test be, and how difficult should my test be?

6. The first step in constructing a test is to answer the question, "Why am I testing?" You need to identify the purpose of the test clearly. What are you trying to find out? What information do you need to get from the results? How will you use this information? In fact, is the test really necessary?

7. Having a clear purpose for a test is particularly important in the classroom situation. Too often, students' time is wasted on tests for which the results are filed away in permanent records without being used for guidance, placement or evaluation.

8. A test or questionnaire for research also needs a clear purpose. Many research reports, for example, merely describe the results obtained when two tests were given to the same group of people.

9. The most common use of a test is to provide information needed to make decisions. Here are some examples of questions that may be answered by test-based information.

---

[1] The numbers alongside the text are used to position the slides that have been developed for the filmstrip.

10. Comparisons of instructional methods or materials. For example, you might want to compare Program A to Program B in terms of student comprehension.

11. Assessment of student learning over a time period: How much more do students know about a particular subject now than they did six weeks ago?

12. Mastery of instructional objectives: Which objectives have been attained by each student in a course? The use of tests to determine mastery has been given many names by test makers and test users. In addition to the term, "mastery testing," the terms, "diagnostic testing" and "criterion referenced testing," have also been applied.

13. Once you decide that you have a clear purpose for a test, it is still possible that you may not have to construct your own. You may be able to use existing instruments that would achieve the purpose as well or better. How do you find out what is available?

14. One of the best current information sources on existing tests is the Mental Measurements Yearbook, a collection of test descriptions and reviews. The Yearbooks are most useful as reference sources when they are combined with copies of the book, Tests in Print, an index to available tests and to reviews in the Yearbooks. Look for recent editions of Tests in Print and the Mental Measurements Yearbook.

15. Another useful publication for information about classroom tests is the CSE Elementary School Test Evaluations. You may also want to check recent professional journals for test reviews and research reports in your area of interest. Much information can also be obtained from the catalogues of Test Publishers.

16. Compare the existing tests with what you would be able to produce yourself. Are they suited for your audience? Do they provide the kind of results you need to make your decisions?

17. Keep in mind that the shortcomings of existing tests may stem from the elusive nature of the attribute being measured -- such as creativity or motivation. These problems could apply equally to your own test.

18. Also, existing tests may have the advantage of careful editing and pretesting which may be too expensive and time consuming for you to carry out for your own test.

19. Another important consideration is the kinds of comparisons needed. If, for example, you want to compare your results to national norms, or to the results of a previous study, it will probably be necessary to use an existing test to make the comparison valid.

20. You will have to develop a new test when you want to measure something not measured adequately in existing tests and when no far-reaching comparisons are required. For example, you would need to develop a new test to evaluate a particular curriculum which approaches content or skills in a unique way not measured by typical survey achievement tests.

21. You would also need a new test if you wanted to assess a small area of course content in great detail. It is possible to use the results from one or two questions on a survey test to make inferences about a student's strengths in an area, but you should not place much confidence in such inferences unless they are supported by additional evidence from further testing, observations, or other evaluations of student performance. In general, a single test should have many questions on a topic if it will be the primary basis for diagnostic judgments about student performance.

22. If you are interested in questionnaires, a careful analysis of available material will often lead you to develop your own instrument. Questionnaires tend to be much more closely tied to a given setting than are tests.

23. Now, assuming that you are going to develop your own test, you need to decide what you should test. A test is by its nature a sample. It consists of a limited number of tasks or questions that represent the total attributes or knowledge that you are trying to assess.

24. One of the primary objectives in setting specifications, therefore, is to make sure your test will contain a representative and balanced selection of tasks or questions.

25. The best way to make sure of this is to use a Content and Skills grid or matrix. It is used to describe what you should test, based on careful consideration of the exact purpose the test is to serve. The first step in developing a Content and Skills grid is to specify the content areas to be included.

26. All appropriate sources for content should be considered. For classroom achievement tests this could include textbooks, teacher's manuals, course syllabi, lesson plans and lecture notes. For a research questionnaire or test, the objectives of the study should be your most useful source for content areas.

27. Your first content list should be as comprehensive as possible, even including categories that are closely related to each other.

28. The next step is to look for meaningful ways to group your categories. Are there any general classifications of content that seem to cut across many of your individual categories?

29. For example, this set of content categories was listed in the preparation of content specifications for a test in American History and Social Studies at the senior high school level.

30. In addition to the content categories, the chronological dimension was judged to be a significant one.

31. These two sets of categories provided the basis for development of a Content and Chronology outline. Identifying the content categories is only the first step in deciding what a test will cover. It is also necessary to decide what proportion of your test will be devoted to each content category.

32. In other words, you must decide how many of your questions will be devoted to each content area. In this sample outline, the relative weight assigned to each area is expressed as a percentage of the total test. You will note that the percentages assigned add up to 100 both horizontally and vertically.

33. The basis for assigning weights should be the relative importance of the material covered. One indicator of importance is the amount of class time and the number of readings devoted to each category during instruction.

34. You will often find it useful to specify the number of questions for each individual content area. This example shows how the questions for the American History and Social Studies test were apportioned. Notice that some aspects of content are more characteristic of one time period than of others.

35. So, full specification of the content area helps you answer the question, "What am I testing?." But there is another dimension to consider for the same question.

36. You must also determine the level of understanding that will be required for answering the questions. Each of the content areas will contain factual material. It will also contain material which could serve as the basis for questions requiring inference and the analysis of relationships.

37. In the American History and Social Studies example, students could be asked questions of fact. The questions could cover points such as the identification of major figures, (Who was the leader of a certain political party), or the dates of events (When did it happen).

38. Or, they could also be asked to test the usefulness of a theory in explaining an event. Essentially, the relative emphasis on recall versus higher level understanding should be the same in a test as it was during instruction.

39. If, for example, only about a third of the measurable objectives of instruction are concerned with the mastery of facts, this area should receive only about a third of the total number of questions in a test.

40. The Taxonomy of Educational Objectives is a useful reference for considering different skills or abilities to be tested.

41. This generalized treatment of the educational process suggests that six major categories of skills can be identified. The categories provided in the Taxonomy of Educational Objectives should help you to decide what skills you are examining. You will, however, have to adapt the approach to your particular subject matter.

42. In the example, six categories were identified.

1. Knowledge of specifics
2. Comprehension of the meaning of various types of historical data.
3. Inferring consequences and conclusions from historical data.
4. Application of abstractions to historical particulars
5. Evaluation of historical data for a given purpose
6. Synthesis of historical data into a new pattern

43. When developing and using specifications of skills, pay close attention to the number of questions involving knowledge or recall and how many demand higher level skills such as perception of relationships.

Make certain that you do not include more questions that demand simple content recall than you specified.

44. It is best to view content and skills grids or matrices as draft documents that may need revision after you carry out the tasks of writing and reviewing questions. You can make some changes to your specifications before you prepare your final test. Besides finding that knowledge questions are easier to write than other types, you will likely discover some content areas which are much richer or more barren of content than you originally estimated.

45. When preparing questionnaires, some areas that seemed highly significant during the planning stages may prove to lack the definition required for successful questioning. Alternatively, it may prove impossible to question the appropriate subjects,so an area of interest may need to be approached from a different perspective.

46. A reviewing process is essential to insure that the final instrument will in fact fit your specifications. It is often a good idea to have someone other than yourself review both the instrument and the specifications. An outsider can point out inconsistencies that you may have overlooked.

47. After you have administered your test or questionnaire, you should review your specifications before preparing any additional instruments. You should, in other words, ask the question, "What should I test?" each time you undertake instrument development.

48. The next question is "whom am I testing?" In the course of making all of your decisions about specifications, you should keep in mind the characteristics of the group to be tested. There are four attributes to consider.

49. The first is age . . . and, in this connection, previous familiarity with tests of various kinds. When dealing with young children you may need a number of practice questions as a warm-up for the real test.

50. The second is language facility. Test and questionnaire developers often include difficult vocabulary and complex sentences in instruments that are to be used with elementary school children. Problems of comprehension can also be substantial when tests or questionnaires are used with non-native speakers. If people don't understand the questions, you won't get very useful answers.

51. Level and distribution of ability in the group that is to be tested. Do you have a group with a typical range and distribution of abilities, or do you have two or more fairly distinct groups?

52. And, finally, you must consider motivation to cooperate. Few people will show much interest in a test unless they have some personal benefit to gain from taking it.

53. Now we come to the important subject of what kinds of questions to use. A number of considerations will influence this decision. Obviously, you will want to use a format that gets at the kinds of knowledge or skills that you are particularly interested in. To some extent, the nature of the subject matter will influence your choices.

54. If you want to know how a student feels about some issue, the use of questions that test his degree of agreement is often the most efficient method.

55. If you are interested in the student's ability to translate from another language into English, the use of material in that language is clearly called for.

56. If you are interested in evaluating a student's map reading skills, the use of maps is indicated.

57. These are some of the question or item types used on tests. While any of these types is appropriate in some circumstances, the essay test tends to be used more often than is warranted. Essays are an extremely inefficient method of obtaining reliable information about individuals. They are only useful as measurement devices when multiple, independent grading is arranged, using graders who can agree in advance on common criteria. Also, the time required to write essays drastically limits your ability to sample many aspects of content.

58. In most testing situations, a balanced blend of several choice response item types will prove most effective. In general, the broader your sampling of content, the more that the scores will be representative of all aspects of content in the area of interest.

59. The next question is "how long should your test be?"

60. This decision often reflects compromises among competing concerns such as time, age, and motivation to cooperate.

61. The number of questions also depends in part on the kinds of questions used. Questions requiring the reading of stimulus material will take much longer than brief independent questions. Questions requiring careful analysis of possibilities will be more time consuming than simple recall questions.

62. It will sometimes be possible to pretest questions to determine the length of time required to answer them as well as to learn about possible question difficulties and ambiguities. Typically, however, you will have to use your own judgment, perhaps using results from similar instruments as a guide.

63. One consideration regarding test length is the degree of reliability or reproduceability of measurement that is needed. The graph shows the relationship between reliability and number of questions. In general, the more questions and the greater their similarity, the more reliable a test will be. But it gets harder and harder to increase reliability by adding items once you reach the upper levels of reliability.

64. The level of reliability needed will depend on the purpose of the test. A test for measuring group performance only, may have much lower reliabilities than a test used to make decisions about or give advice to individuals.

65. Reversability of decisions is also important. If decisions based on test scores, such as grouping students for a particular unit of instruction, can be revised if they prove incorrect or harmful, less precision of measurement is necessary. If, on the other hand, an important and hard to reverse decision is to be made on the basis of test scores, high reliability of measurement is essential. An example of such a decision would be that of placing a student in a class for the mentally retarded.

66. If the test depends on observing the subject's reactions, individual testing rather than group testing will be necessary and the time to administer each test becomes important. Wherever possible, group testing is preferable to individual testing from a time point of view.

67. Next comes the vital consideration of how difficult your test should be.

68. The difficulty of a test is also influenced by the purposes of testing. If you are conducting an assessment of a large block of content and are equally interested in the performance displayed by students at each level of proficiency, you will be looking for a spread of scores similar to that

shown in this graph. To achieve this result, you will want a test of middle difficulty with items ranging from below middle difficulty to above it.

69. If you are primarily interested in differentiating your lowest scoring students from other students, you would be aiming for the spread of scores shown here. This could be the case when selecting students for remediation. To achieve this, you would select many of your items to be easy for the total group. This would give you the necessary differentiation among your low scoring students even though it would not differentiate effectively among high scoring students where scores tend to be bunched together.

70. If your purpose for testing is to see whether certain fundamental facts or principles have been acquired, you will want to examine the performance of students on each question. You may not be particularly interested in what the total score distribution looks like.

71. When you are concerned about score distributions as would be the case when you were assigning grades to students or ranking individuals for some purpose you must consider the spread of difficulty among the individual questions as well as the overall level of difficulty of a test. In this connection, remember that a question of middle difficulty provides you with the maximum amount of information that can be obtained from a single question.

72. Assume that you have 100 students responding to a question which can be answered right or wrong. If the question is answered correctly by 50 students this differentiates each of them from each of the 50 students who answered the question incorrectly. Thus, a middle difficulty question helps you make 50 times 50 or 2,500 differentiations.

73. But a question that is answered correctly by as few as 10 or as many as 90 of the 100 students helps you make only 10 times 90 or 900 differentiations. Thus, the closer the question is to middle difficulty, the more useful the results can be.

74. One useful approach to controlling the difficulty of a test is to aim for proportions of questions at various difficulty levels such as 1/3 easy, 1/3 medium and 1/3 hard. You may find that easy questions or, less often, hard questions are the most difficult to write.

75. You may wonder why all questions in a test should not be of middle difficulty. The answer is that aspects of content vary in difficulty. If you sample a large area of content adequately, you will get some relatively easy and some relatively hard content areas.

76. You might set as a target 50% easy, 30% medium and 20% hard questions. Your easy and hard questions should, of course, remain close enough to middle difficulty to provide good differentiation among students. It will be necessary to cope with a strong tendency to overestimate what students know. If you can pretest your questions, this problem of estimating can be made very apparent.

77. A general rule to keep in mind is that a middle difficulty or near middle difficulty test is often preferable even when your primary interest is in groups of students scoring substantially above or below average.

78. To summarize, these six questions need to be answered in the course of planning a test:

Why am I testing?
What should I test?
Whom am I testing?
What kinds of questions should I use?
How long should my test be?
How difficult should my test be?

79. Some of the major points that have been made are the following:

Tests should only be developed or selected and administered in situations where there is a clear and useful purpose for the tests.

80. There are many situations in which it will be appropriate to use existing tests, rather than to develop new tests, particularly where comparisons with other groups are needed.

81. If you develop your own test you will find that a two dimensional Content and Skills grid or matrix is a valuable way to insure that your test will contain a representative and balanced selection of tasks or questions.

82. When preparing a test you need to keep in mind the nature of the group being tested, including such attributes as:

--age
--language facility
--level and distribution of ability in the group
--motivation to cooperate

83. In most testing settings you should consider the use of a balanced blend of choice response questions that cover a broad area of content and skills.

84. In deciding on the number of questions to be included in your test, you will usually need to review the following issues:

--the kinds of questions you will use
--practical constraints such as length of a classroom period, and
--required reliability, and in that connection, importance and reversability of decisions.

85. And finally, the level of difficulty of your test should be decided by the purpose for which the test is to be used.

86. Office of Education credit.

87. J. Fremer & other ETS staff credits.

88. Visual Education Corporation credits.

89. Visual Education Corporation credits.

STATE UNIVERSITY OF NEW YORK
UPSTATE MEDICAL CENTER
766 IRVING AVENUE
SYRACUSE, N. Y. 13210

EDUCATIONAL COMMUNICATIONS

AREA CODE 315
473-4860

July 20, 1973

Dr. John Fremer
Educational Testing Service
Princeton, New Jersey 08540

Dear John:

After N years, I finally used the tape on test construction that you had sent to me. I showed it to two fellows here in the medical school who are developing some cognitive instruments for courses they are teaching.

I need to tell you that they are wildly enthusiastic and they suggested the medical school get its own copy of the film strip for use here. My personal feeling is that it should be submitted to the Canne Film Festival.

Thanks again for sending me the materials.

Cordially,

Hank

Henry Slotnick, Ph.D.
Assistant Professor

HS:aml

STATE UNIVERSITY OF NEW YORK
UPSTATE MEDICAL CENTER
766 IRVING AVENUE
SYRACUSE, N. Y. 13210

EDUCATIONAL COMMUNICATIONS

AREA CODE 315
473-4860

July 20, 1973

Dr. John Fremer
Educational Testing Service
Princeton, New Jersey 08540

Dear Dr. Fremer:

I previewed your filmstrip program entitled "Planning a Test", HEW OEC-0-70-4777 and I would like to purchase a copy of these materials preferably in cassette-slide format.

Please let me know the cost and I will be glad to send a purchase requisition.

Sincerely,

James S. Waldron, Ph.D.
Director

JSW:aml

Oral Roberts University
7777 South Lewis • Tulsa, Oklahoma 74105

Rec'd at RTO -
SEP 27 1973

OFFICE OF THE VICE-PRESIDENT
FOR LEARNING RESOURCES AND INSTRUCTION

September 19, 1973

Dr. Donald E. Hood
Educational Testing Service
The Quadrangle, Suite 253
3810 Medical Parkway
Austin, Texas 78756

Dear Dr. Hood:

I am enclosing the tape and filmstrip that you used while here for the testing workshop. Comments from the faculty about the workshop have been all positive and more than that, the chairmen and the faculty are getting more involved in a real testing program. I am endeavoring to work with the chairmen in their follow-up work with the faculty on their testing methods and levels.

Don, thanks for your help and please keep me apprised of developments at Educational Testing Service that you feel would benefit the ORU program. You have left a very positive image at ORU for yourself and Educational Testing Service, and as you already know, you are welcome here at ORU.

Sincerely,

William W. Jernigan, Vice-President
Learning Resources and Instruction

WWJ/sw

kansas state teache[r] [c]ollege
Data Processing and Educational Measurements Center
[C]OMMERCIAL STREET
[E]MIA, KANSAS 66801
[T]ELEPHONE 316 343-1200

October 18, 1973

Dr. John Fremer
Associate Director, Elementary and
Secondary School Programs
Educational Testing Service
Princeton, N. J. 08540

Dear John,

The filmstrip is excellent, it was received with "excessive enthusiasm" by two classes of graduate students in measurement..

Let me know when they will be available for purchase as we would like to buy one.

Take care.

Sincerely,

Howard P. Schwartz

Howard P. Schwartz, Ed.D.
Bureau of Educational Measurements

HPS/mar